Measuring implications of the *D*-basis in biomedical applications

K. Adaricheva

Nazarbayev University (NU), Astana

ICFCA-2015, June 26
Nerja, Malaga

Based on papers:

“Ordered direct basis of a finite closure system"
joint work with J.B.Nation and R.Rand, Disc. Appl. Math. 2013

“Discovery of the D-basis in binary tables based on hypergraph dualization", 2012
joint work with J.B.Nation, submitted, in arxiv

“Measuring the implications of the D-basis in analysis of data in biomedical studies", 2015
joint work with J.B.Nation, G. Okimoto, K. Alibek and others, Proceedings of ICFCA-2015, Spain

Forthcoming publication

KA and J.B.Nation, four chapters for:
Lattice Theory: Special Topics and Applications, volume 2,
G. Grätzer and F. Wehrung, eds., Springer, to appear in 2015

Last paper's support:

The project has been supported by the research grant N 13/42
“ Algebraic methods of data retrieval"
Nazarbayev University, 2013-2015
and grant N 0112PK02175, 2012-2014,
Ministry of Healthcare and Social Development of RK

Acknowledgments

The following people assisted in the project:

- Joshua Blumenkopf (Yeshiva College, New York, 3d year Physics major in 2013)
- Toviah Moldwin (Yeshiva College, New York, 3d year CS major in 2013)
- Takeaki Uno (National Institute of Informatics, Tokyo)
- Gordon Okimoto (University of Hawaii Cancer Center)
- Nazar Seidalin (hospital of Medical Holding, Astana)
- Kenneth Alibek (Graduate School of Medicine, NU)
- Vyacheslav Adarichev (Biology Department, NU)
- Adina Amanbekkyzy (Math Department TA, NU)
- Shuchismita Sarkar (Math Department TA, NU)
- Alibek Sailanbayev (2d year CS student, NU)
- Ulrich Norbisrath (CS Department, NU)
- Mark Sterling (CS Department, NU)

Outline

Informally

We will see

- a basis shorter than shortest direct basis, but still possessing directness property;
- algorithm that requires knowledge and understanding of (concept) lattices, but does not need the (concept) lattice itself;
- an idea how to compute association rules using implications;
- an approach to rank implications which did not exist before.

Informally

We will see

- a basis shorter than shortest direct basis, but still possessing directness property;
- algorithm that requires knowledge and understanding of (concept) lattices, but does not need the (concept) lattice itself;
- an idea how to compute association rules using implications;
- an approach to rank implications which did not exist before.

Informally

We will see

- a basis shorter than shortest direct basis, but still possessing directness property;
- algorithm that requires knowledge and understanding of (concept) lattices, but does not need the (concept) lattice itself;
- an idea how to compute association rules using implications;
- an approach to rank implications which did not exist before.

Informally

We will see

- a basis shorter than shortest direct basis, but still possessing directness property;
- algorithm that requires knowledge and understanding of (concept) lattices, but does not need the (concept) lattice itself;
- an idea how to compute association rules using implications;
- an approach to rank implications which did not exist before.

Closure systems

$\langle X, \phi \rangle$ is *a closure system*, if

- X is non-empty set (finite in this talk);
- ϕ is a closure operator on X, i.e. $\phi : 2^X \to 2^X$ with
  (1) $Y \subseteq \phi(Y)$;
  (2) $Y \subseteq Z$ implies $\phi(Y) \subseteq \phi(Z)$;
  (3) $\phi(\phi(Y)) = \phi(Y)$, for all $Y, Z \subseteq X$.
- Closed set: $A = \phi(A)$;
- Lattice of closed sets: $Cl(X, \phi)$.

Closure systems

$\langle X, \phi \rangle$ is *a closure system*, if

- X is non-empty set (finite in this talk);
- ϕ is a closure operator on X, i.e. $\phi : 2^X \to 2^X$ with
  - (1) $Y \subseteq \phi(Y)$;
  - (2) $Y \subseteq Z$ implies $\phi(Y) \subseteq \phi(Z)$;
  - (3) $\phi(\phi(Y)) = \phi(Y)$, for all $Y, Z \subseteq X$.
- Closed set: $A = \phi(A)$;
- Lattice of closed sets: $Cl(X, \phi)$.

Closure systems

$\langle X, \phi \rangle$ is *a closure system*, if

- X is non-empty set (finite in this talk);
- ϕ is a closure operator on X, i.e. $\phi : 2^X \to 2^X$ with
  - (1) $Y \subseteq \phi(Y)$;
  - (2) $Y \subseteq Z$ implies $\phi(Y) \subseteq \phi(Z)$;
  - (3) $\phi(\phi(Y)) = \phi(Y)$, for all $Y, Z \subseteq X$.
- Closed set: $A = \phi(A)$;
- Lattice of closed sets: $Cl(X, \phi)$.

Lattices and closure systems

Proposition

Every finite lattice L is the lattice of closed sets of some closure system $\langle X, \phi \rangle$.

- Take $X = \mathrm{Ji}(L)$, the set of join-irreducible elements: $j \in \mathrm{Ji}(L)$, if $j \neq 0$, and $j = a \vee b$ implies $j = a$ or $j = b$;
- Define $\phi(Y) = \{j \in \mathrm{Ji}(L) : j \leq \bigvee Y\}$, $Y \subseteq X$.
- Proof of $L \cong Cl(X, \phi)$: every element $x \in L$ corresponds to ϕ-closed set of all join irreducibles below x.
- So defined closure system is always *standard*.

Lattices and closure systems

Proposition

Every finite lattice L is the lattice of closed sets of some closure system $\langle X, \phi \rangle$.

- Take $X = \mathrm{Ji}(L)$, the set of join-irreducible elements: $j \in \mathrm{Ji}(L)$, if $j \neq 0$, and $j = a \vee b$ implies $j = a$ or $j = b$;
- Define $\phi(Y) = \{j \in \mathrm{Ji}(L) : j \leq \bigvee Y\}$, $Y \subseteq X$.
- Proof of $L \cong Cl(X, \phi)$: every element $x \in L$ corresponds to ϕ-closed set of all join irreducibles below x.
- So defined closure system is always *standard*.

Lattices and closure systems

Proposition

Every finite lattice L is the lattice of closed sets of some closure system $\langle X, \phi \rangle$.

- Take $X = \mathrm{Ji}(L)$, the set of join-irreducible elements: $j \in \mathrm{Ji}(L)$, if $j \neq 0$, and $j = a \vee b$ implies $j = a$ or $j = b$;
- Define $\phi(Y) = \{j \in \mathrm{Ji}(L) : j \leq \bigvee Y\}$, $Y \subseteq X$.
- Proof of $L \cong Cl(X, \phi)$: every element $x \in L$ corresponds to ϕ-closed set of all join irreducibles below x.
- So defined closure system is always *standard*.

Lattices and closure systems

Proposition

Every finite lattice L is the lattice of closed sets of some closure system $\langle X, \phi \rangle$.

- Take $X = \mathrm{Ji}(L)$, the set of join-irreducible elements: $j \in \mathrm{Ji}(L)$, if $j \neq 0$, and $j = a \vee b$ implies $j = a$ or $j = b$;
- Define $\phi(Y) = \{j \in \mathrm{Ji}(L) : j \leq \bigvee Y\}$, $Y \subseteq X$.
- Proof of $L \cong Cl(X, \phi)$: every element $x \in L$ corresponds to ϕ-closed set of all join irreducibles below x.
- So defined closure system is always *standard*.

Lattices and closure systems

Proposition

Every finite lattice L is the lattice of closed sets of some closure system $\langle X, \phi \rangle$.

- Take $X = \mathrm{Ji}(L)$, the set of join-irreducible elements: $j \in \mathrm{Ji}(L)$, if $j \neq 0$, and $j = a \vee b$ implies $j = a$ or $j = b$;
- Define $\phi(Y) = \{j \in \mathrm{Ji}(L) : j \leq \bigvee Y\}$, $Y \subseteq X$.
- Proof of $L \cong Cl(X, \phi)$: every element $x \in L$ corresponds to ϕ-closed set of all join irreducibles below x.
- So defined closure system is always *standard*.

Lattices and closure systems

Proposition

Every finite lattice L is the lattice of closed sets of some closure system $\langle X, \phi \rangle$.

- Take $X = \mathrm{Ji}(L)$, the set of join-irreducible elements: $j \in \mathrm{Ji}(L)$, if $j \neq 0$, and $j = a \vee b$ implies $j = a$ or $j = b$;
- Define $\phi(Y) = \{j \in \mathrm{Ji}(L) : j \leq \bigvee Y\}$, $Y \subseteq X$.
- Proof of $L \cong Cl(X, \phi)$: every element $x \in L$ corresponds to ϕ-closed set of all join irreducibles below x.
- So defined closure system is always *standard*.

Standard closure systems

- Closure system $\langle X, \phi \rangle$ is standard, when closed set $\phi(\{x\})$, for every $x \in X$, has a unique minimal generator $\{x\}$.
- In the standard system $\langle X, \phi \rangle$, closed sets $\phi(\{x\})$, $x \in X$, are exactly join-irreducibles of lattice $Cl(X, \phi)$.
- Among all closure systems with the same closure lattice, a standard one is defined on the smallest base set.
- Every closure system $\langle Y, \phi \rangle$ can be reduced to a standard one on some $X \subseteq Y$.

Standard closure systems

- Closure system $\langle X, \phi \rangle$ is standard, when closed set $\phi(\{x\})$, for every $x \in X$, has a unique minimal generator $\{x\}$.
- In the standard system $\langle X, \phi \rangle$, closed sets $\phi(\{x\})$, $x \in X$, are exactly join-irreducibles of lattice $Cl(X, \phi)$.
- Among all closure systems with the same closure lattice, a standard one is defined on the smallest base set.
- Every closure system $\langle Y, \phi \rangle$ can be reduced to a standard one on some $X \subseteq Y$.

Standard closure systems

- Closure system $\langle X, \phi \rangle$ is standard, when closed set $\phi(\{x\})$, for every $x \in X$, has a unique minimal generator $\{x\}$.
- In the standard system $\langle X, \phi \rangle$, closed sets $\phi(\{x\})$, $x \in X$, are exactly join-irreducibles of lattice $Cl(X, \phi)$.
- Among all closure systems with the same closure lattice, a standard one is defined on the smallest base set.
- Every closure system $\langle Y, \phi \rangle$ can be reduced to a standard one on some $X \subseteq Y$.

Standard closure systems

- Closure system $\langle X, \phi \rangle$ is standard, when closed set $\phi(\{x\})$, for every $x \in X$, has a unique minimal generator $\{x\}$.
- In the standard system $\langle X, \phi \rangle$, closed sets $\phi(\{x\})$, $x \in X$, are exactly join-irreducibles of lattice $Cl(X, \phi)$.
- Among all closure systems with the same closure lattice, a standard one is defined on the smallest base set.
- Every closure system $\langle Y, \phi \rangle$ can be reduced to a standard one on some $X \subseteq Y$.

Example: Building a closure system associated with lattice A_{12}.
$X = \mathrm{Ji}(A_{12}) = \{1, 2, 3, 4, 5, 6\}$. $\phi(\{4, 6\}) = \{1, 3, 4, 6\}$, $\phi(\{2, 4\}) = X$ etc.

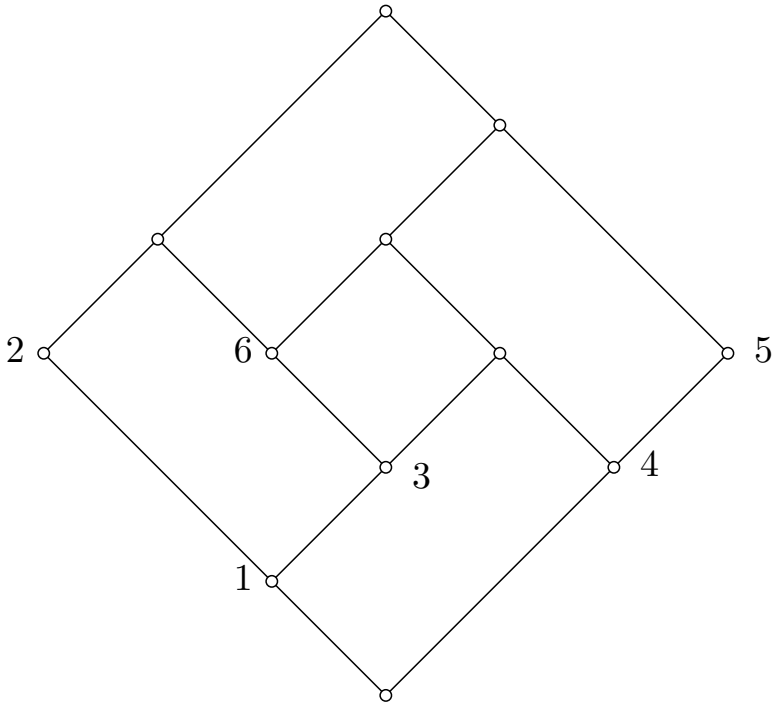


Figure: A_{12}

Closure systems and implications

- An implication σ on X: $Y \to Z$, for $Y, Z \subseteq X$, $Z \neq \emptyset$.
- σ-closed subset A of X: if $Y \subseteq A$, then $Z \subseteq A$.
- Closure system $\langle X, \phi_{\mathcal{S}} \rangle$ defined by set $\mathcal{S}$ of implications on X: A is closed, if it is σ-closed, for each $\sigma \in \mathcal{S}$
- Every set of implications $\mathcal{S}$ on X defines *unique* closure operator on X.

Closure systems and implications

- An implication σ on X: $Y \to Z$, for $Y, Z \subseteq X$, $Z \neq \emptyset$.
- σ-closed subset A of X: if $Y \subseteq A$, then $Z \subseteq A$.
- Closure system $\langle X, \phi_{\mathcal{S}} \rangle$ defined by set $\mathcal{S}$ of implications on X: A is closed, if it is σ-closed, for each $\sigma \in \mathcal{S}$
- Every set of implications $\mathcal{S}$ on X defines *unique* closure operator on X.

Closure systems and implications

- An implication σ on X: $Y \to Z$, for $Y, Z \subseteq X$, $Z \neq \emptyset$.
- σ-closed subset A of X: if $Y \subseteq A$, then $Z \subseteq A$.
- Closure system $\langle X, \phi_{\mathcal{S}} \rangle$ defined by set $\mathcal{S}$ of implications on X: A is closed, if it is σ-closed, for each $\sigma \in \mathcal{S}$
- Every set of implications $\mathcal{S}$ on X defines *unique* closure operator on X.

Closure systems and implications

- An implication σ on X: $Y \to Z$, for $Y, Z \subseteq X$, $Z \neq \emptyset$.
- σ-closed subset A of X: if $Y \subseteq A$, then $Z \subseteq A$.
- Closure system $\langle X, \phi_{\mathcal{S}} \rangle$ defined by set $\mathcal{S}$ of implications on X: A is closed, if it is σ-closed, for each $\sigma \in \mathcal{S}$
- Every set of implications $\mathcal{S}$ on X defines *unique* closure operator on X.

Transformation of an implication to the Horn Formula

- *Unit implication* σ on X: $Y \to z,\ Y \subseteq X, z \in X$.
- Every implication $Y \to Z$ is equivalent to the set of unit implications $\{Y \to z, z \in Z\}$: *unit expansion*.
- Logical interpretation of unit implication σ:
  $X = \{x_1, \ldots, x_n\}, Y = \{x_1, \ldots, x_k\}, z = x_{k+1}$
  $\sigma \equiv (x_1 \wedge x_2 \cdots \wedge x_k \to x_{k+1}) \equiv (\neg x_1 \vee \neg x_2 \vee \cdots \vee \neg x_k \vee x_{k+1})$.

Transformation of an implication to the Horn Formula

- *Unit implication* σ on X: $Y \to z,\ Y \subseteq X, z \in X$.
- Every implication $Y \to Z$ is equivalent to the set of unit implications $\{Y \to z, z \in Z\}$: *unit expansion*.
- Logical interpretation of unit implication σ:
  $X = \{x_1, \ldots, x_n\}, Y = \{x_1, \ldots, x_k\}, z = x_{k+1}$
  $\sigma \equiv (x_1 \wedge x_2 \cdots \wedge x_k \to x_{k+1}) \equiv (\neg x_1 \vee \neg x_2 \vee \cdots \vee \neg x_k \vee x_{k+1})$.

Transformation of an implication to the Horn Formula

- *Unit implication* σ on X: $Y \to z,\ Y \subseteq X, z \in X$.
- Every implication $Y \to Z$ is equivalent to the set of unit implications $\{Y \to z, z \in Z\}$: *unit expansion*.
- Logical interpretation of unit implication σ:
  $X = \{x_1, \ldots, x_n\}$, $Y = \{x_1, \ldots, x_k\}$, $z = x_{k+1}$
  $\sigma \equiv (x_1 \wedge x_2 \cdots \wedge x_k \to x_{k+1}) \equiv (\neg x_1 \vee \neg x_2 \vee \cdots \vee \neg x_k \vee x_{k+1})$.

Many faces of implications

An implication may appear as:

- Ordered pair of subsets of base set X.
- Propositional Horn formula on the set of variables X.
- CNF of Horn Boolean function on variables X.
- An edge of directed hypergraph on set of vertices X.

Closure systems and implications

- Every closure system $\langle X, \psi \rangle$ can be presented as $\langle X, \phi_{\mathcal{S}} \rangle$, for some set $\mathcal{S}$ of implications on X.
- Example: $\mathcal{S} = \{A \to \psi(A) : A \subseteq X, A \neq \psi(A)\}$.
- There exist *numerous* sets of implications that define the same operator on X.
- Term *a base* or *a basis* is used when the set of implications $\mathcal{S}'$ that defines the same closure system satisfies some condition of minimality.
- Two famous bases: Guigues-Duquenne and canonical direct unit basis.

Closure systems and implications

- Every closure system $\langle X, \psi \rangle$ can be presented as $\langle X, \phi_{\mathcal{S}} \rangle$, for some set $\mathcal{S}$ of implications on X.
- Example: $\mathcal{S} = \{A \to \psi(A) : A \subseteq X, A \neq \psi(A)\}$.
- There exist *numerous* sets of implications that define the same operator on X.
- Term *a base* or *a basis* is used when the set of implications $\mathcal{S}'$ that defines the same closure system satisfies some condition of minimality.
- Two famous bases: Guigues-Duquenne and canonical direct unit basis.

Closure systems and implications

- Every closure system $\langle X, \psi \rangle$ can be presented as $\langle X, \phi_{\mathcal{S}} \rangle$, for some set $\mathcal{S}$ of implications on X.
- Example: $\mathcal{S} = \{A \to \psi(A) : A \subseteq X, A \neq \psi(A)\}$.
- There exist *numerous* sets of implications that define the same operator on X.
- Term *a base* or *a basis* is used when the set of implications $\mathcal{S}'$ that defines the same closure system satisfies some condition of minimality.
- Two famous bases: Guigues-Duquenne and canonical direct unit basis.

Closure systems and implications

- Every closure system $\langle X, \psi \rangle$ can be presented as $\langle X, \phi_{\mathcal{S}} \rangle$, for some set $\mathcal{S}$ of implications on X.
- Example: $\mathcal{S} = \{A \to \psi(A) : A \subseteq X, A \neq \psi(A)\}$.
- There exist *numerous* sets of implications that define the same operator on X.
- Term *a base* or *a basis* is used when the set of implications $\mathcal{S}'$ that defines the same closure system satisfies some condition of minimality.
- Two famous bases: Guigues-Duquenne and canonical direct unit basis.

Closure systems and implications

- Every closure system $\langle X, \psi \rangle$ can be presented as $\langle X, \phi_{\mathcal{S}} \rangle$, for some set $\mathcal{S}$ of implications on X.
- Example: $\mathcal{S} = \{A \to \psi(A) : A \subseteq X, A \neq \psi(A)\}$.
- There exist *numerous* sets of implications that define the same operator on X.
- Term *a base* or *a basis* is used when the set of implications $\mathcal{S}'$ that defines the same closure system satisfies some condition of minimality.
- Two famous bases: Guigues-Duquenne and canonical direct unit basis.

Guigues-Duquenne basis

D. Maier, *Minimum covers in the relational database model*, JACM **27** (1980), 664–674.
J.L. Guigues, V. Duquenne, *Familles minimales d'implications informatives résultant d'une tables de données binares*, Math. Sci. Hum. **95** (1986), 5–18.

- Define *critical subsets of* X for a given closure system $\langle X, \phi \rangle$.
- Canonical basis $\mathcal{S}_C$ is $\{C \to B : C \text{ is critical}, B = \phi(C) \setminus C\}$.
- If $\mathcal{S}$ is any other set of implications generating $\langle X, \phi \rangle$, then for every critical set C one can find $(C' \to D) \in \mathcal{S}$ such that $C' \subseteq C$, and no other critical or closed set Y with $C' \subseteq Y \subset C$.
- $\mathcal{S}_C$ is the *minimum* basis among all the bases generating $\langle X, \phi \rangle$.

Definition

Set of implications $\mathcal{S}$ defining ϕ on A is called minimum basis if $|\mathcal{S}|$ is minimal among all sets of implications defining ϕ.

Guigues-Duquenne basis

D. Maier, *Minimum covers in the relational database model*, JACM **27** (1980), 664–674.
J.L. Guigues, V. Duquenne, *Familles minimales d'implications informatives résultant d'une tables de données binares*, Math. Sci. Hum. **95** (1986), 5–18.

- Define *critical subsets of X* for a given closure system $\langle X, \phi \rangle$.
- Canonical basis $\mathcal{S}_C$ is $\{C \to B : C \text{ is critical}, B = \phi(C) \setminus C\}$.
- If $\mathcal{S}$ is any other set of implications generating $\langle X, \phi \rangle$, then for every critical set C one can find $(C' \to D) \in \mathcal{S}$ such that $C' \subseteq C$, and no other critical or closed set Y with $C' \subseteq Y \subset C$.
- $\mathcal{S}_C$ is the *minimum* basis among all the bases generating $\langle X, \phi \rangle$.

Definition

Set of implications $\mathcal{S}$ defining ϕ on A is called minimum basis if $|\mathcal{S}|$ is minimal among all sets of implications defining ϕ.

Guigues-Duquenne basis

D. Maier, *Minimum covers in the relational database model*, JACM **27** (1980), 664–674.
J.L. Guigues, V. Duquenne, *Familles minimales d'implications informatives résultant d'une tables de données binares*, Math. Sci. Hum. **95** (1986), 5–18.

- Define *critical subsets of X* for a given closure system $\langle X, \phi \rangle$.
- Canonical basis $\mathcal{S}_C$ is $\{C \to B : C \text{ is critical}, B = \phi(C) \setminus C\}$.
- If $\mathcal{S}$ is any other set of implications generating $\langle X, \phi \rangle$, then for every critical set C one can find $(C' \to D) \in \mathcal{S}$ such that $C' \subseteq C$, and no other critical or closed set Y with $C' \subseteq Y \subset C$.
- $\mathcal{S}_C$ is the *minimum* basis among all the bases generating $\langle X, \phi \rangle$.

Definition

Set of implications $\mathcal{S}$ defining ϕ on A is called minimum basis if $|\mathcal{S}|$ is minimal among all sets of implications defining ϕ.

Guigues-Duquenne basis

D. Maier, *Minimum covers in the relational database model*, JACM **27** (1980), 664–674.
J.L. Guigues, V. Duquenne, *Familles minimales d'implications informatives résultant d'une tables de données binares*, Math. Sci. Hum. **95** (1986), 5–18.

- Define *critical subsets of X* for a given closure system $\langle X, \phi \rangle$.
- Canonical basis $\mathcal{S}_C$ is $\{C \to B : C \text{ is critical}, B = \phi(C) \setminus C\}$.
- If $\mathcal{S}$ is any other set of implications generating $\langle X, \phi \rangle$, then for every critical set C one can find $(C' \to D) \in \mathcal{S}$ such that $C' \subseteq C$, and no other critical or closed set Y with $C' \subseteq Y \subset C$.
- $\mathcal{S}_C$ is the *minimum* basis among all the bases generating $\langle X, \phi \rangle$.

Definition

Set of implications $\mathcal{S}$ defining ϕ on A is called minimum basis if $|\mathcal{S}|$ is minimal among all sets of implications defining ϕ.

Guigues-Duquenne basis

D. Maier, *Minimum covers in the relational database model*, JACM **27** (1980), 664–674.
J.L. Guigues, V. Duquenne, *Familles minimales d'implications informatives résultant d'une tables de données binares*, Math. Sci. Hum. **95** (1986), 5–18.

- Define *critical subsets of X* for a given closure system $\langle X, \phi \rangle$.
- Canonical basis $\mathcal{S}_C$ is $\{C \to B : C \text{ is critical}, B = \phi(C) \setminus C\}$.
- If $\mathcal{S}$ is any other set of implications generating $\langle X, \phi \rangle$, then for every critical set C one can find $(C' \to D) \in \mathcal{S}$ such that $C' \subseteq C$, and no other critical or closed set Y with $C' \subseteq Y \subset C$.
- $\mathcal{S}_C$ is the *minimum* basis among all the bases generating $\langle X, \phi \rangle$.

Definition

Set of implications $\mathcal{S}$ defining ϕ on A is called minimum basis if $|\mathcal{S}|$ is minimal among all sets of implications defining ϕ.

Guigues-Duquenne basis

D. Maier, *Minimum covers in the relational database model*, JACM **27** (1980), 664–674.
J.L. Guigues, V. Duquenne, *Familles minimales d'implications informatives résultant d'une tables de données binares*, Math. Sci. Hum. **95** (1986), 5–18.

- Define *critical subsets of X* for a given closure system $\langle X, \phi \rangle$.
- Canonical basis $\mathcal{S}_C$ is $\{C \to B : C \text{ is critical}, B = \phi(C) \setminus C\}$.
- If $\mathcal{S}$ is any other set of implications generating $\langle X, \phi \rangle$, then for every critical set C one can find $(C' \to D) \in \mathcal{S}$ such that $C' \subseteq C$, and no other critical or closed set Y with $C' \subseteq Y \subset C$.
- $\mathcal{S}_C$ is the *minimum* basis among all the bases generating $\langle X, \phi \rangle$.

Definition

Set of implications $\mathcal{S}$ defining ϕ on A is called minimum basis if $|\mathcal{S}|$ is minimal among all sets of implications defining ϕ.

Canonical direct unit basis

K.Bertet, B.Monjardet, *The multiple facets of the canonical direct unit implicational basis*, Theoretical Computer Science 411 (2010), 2155-2166.

Given task of computation a closure of some $Y \subset X$ using (unit) basis $\mathcal{S}$, define
$\pi_{\mathcal{S}}(Y) = Y \cup \bigcup\{b : (A \to b) \in \mathcal{S}, A \subseteq Y\}$.
Then $\phi_{\mathcal{S}}(Y) = \pi_{\mathcal{S}}(Y) \cup \pi^2_{\mathcal{S}}(Y) \cup \pi^3_{\mathcal{S}}(Y) \cup \ldots$

A unit implicational basis is called *direct*, if
$\phi_{\mathcal{S}}(Y) = \pi_{\mathcal{S}}(Y)$, for all $Y \subseteq X$.

Canonical direct unit basis

K.Bertet, B.Monjardet, *The multiple facets of the canonical direct unit implicational basis*, Theoretical Computer Science 411 (2010), 2155-2166.

Given task of computation a closure of some $Y \subset X$ using (unit) basis $\mathcal{S}$, define
$\pi_{\mathcal{S}}(Y) = Y \cup \bigcup\{b : (A \to b) \in \mathcal{S}, A \subseteq Y\}$.
Then $\phi_{\mathcal{S}}(Y) = \pi_{\mathcal{S}}(Y) \cup \pi^2_{\mathcal{S}}(Y) \cup \pi^3_{\mathcal{S}}(Y) \cup \ldots$

A unit implicational basis is called *direct*, if
$\phi_{\mathcal{S}}(Y) = \pi_{\mathcal{S}}(Y)$, for all $Y \subseteq X$.

Canonical direct unit basis

K.Bertet, B.Monjardet, *The multiple facets of the canonical direct unit implicational basis*, Theoretical Computer Science 411 (2010), 2155-2166.

Given task of computation a closure of some $Y \subset X$ using (unit) basis $\mathcal{S}$, define
$\pi_{\mathcal{S}}(Y) = Y \cup \bigcup\{b : (A \to b) \in \mathcal{S}, A \subseteq Y\}$.
Then $\phi_{\mathcal{S}}(Y) = \pi_{\mathcal{S}}(Y) \cup \pi^2_{\mathcal{S}}(Y) \cup \pi^3_{\mathcal{S}}(Y) \cup \ldots$

A unit implicational basis is called *direct*, if
$\phi_{\mathcal{S}}(Y) = \pi_{\mathcal{S}}(Y)$, for all $Y \subseteq X$.

Canonical direct unit basis

K.Bertet, B.Monjardet, *The multiple facets of the canonical direct unit implicational basis*, Theoretical Computer Science 411 (2010), 2155-2166.

Given task of computation a closure of some $Y \subset X$ using (unit) basis $\mathcal{S}$, define
$\pi_{\mathcal{S}}(Y) = Y \cup \bigcup\{b : (A \to b) \in \mathcal{S}, A \subseteq Y\}$.
Then $\phi_{\mathcal{S}}(Y) = \pi_{\mathcal{S}}(Y) \cup \pi^2_{\mathcal{S}}(Y) \cup \pi^3_{\mathcal{S}}(Y) \cup \ldots$

A unit implicational basis is called *direct*, if
$\phi_{\mathcal{S}}(Y) = \pi_{\mathcal{S}}(Y)$, for all $Y \subseteq X$.

Minimality

Theorem (Bertet-Monjardet 2010)

For every finite closure system $\langle X, \phi \rangle$ 5 different bases (introduced independently in the literature between 1983 and 2004) are the same.

This basis is called *canonical unit direct*.

Corollary (Bertet-Monjardet 2010)

For a given closure system $\langle X, \phi \rangle$, the canonical unit direct basis is the least basis, with respect to inclusion, among all unit direct sets of implications defining the system.

Minimality

Theorem (Bertet-Monjardet 2010)

For every finite closure system $\langle X, \phi \rangle$ 5 different bases (introduced independently in the literature between 1983 and 2004) are the same.

This basis is called *canonical unit direct*.

Corollary (Bertet-Monjardet 2010)

For a given closure system $\langle X, \phi \rangle$, the canonical unit direct basis is the least basis, with respect to inclusion, among all unit direct sets of implications defining the system.

Minimality

Theorem (Bertet-Monjardet 2010)

For every finite closure system $\langle X, \phi \rangle$ 5 different bases (introduced independently in the literature between 1983 and 2004) are the same.

This basis is called *canonical unit direct*.

Corollary (Bertet-Monjardet 2010)

For a given closure system $\langle X, \phi \rangle$, the canonical unit direct basis is the least basis, with respect to inclusion, among all unit direct sets of implications defining the system.

Pre-cursor of the *D*-basis

OD-graph of a finite lattice:
J.B.Nation, *An approach to lattice varieties of finite height*,
Alg. Universalis **27** (1990), 521–543.

The full information about finite lattice L can be compactly recorded in

- partially ordered set of join-irreducible elements $\langle \mathrm{Ji}(L), \leq \rangle$;
- the minimal join-covers of join-irreducible elements.

Pre-cursor of the *D*-basis

OD-graph of a finite lattice:
J.B.Nation, *An approach to lattice varieties of finite height*,
Alg. Universalis **27** (1990), 521–543.

The full information about finite lattice L can be compactly recorded in

- partially ordered set of join-irreducible elements $\langle \mathrm{Ji}(L), \leq \rangle$;
- the minimal join-covers of join-irreducible elements.

Pre-cursor of the *D*-basis

OD-graph of a finite lattice:
J.B.Nation, *An approach to lattice varieties of finite height*,
Alg. Universalis **27** (1990), 521–543.

The full information about finite lattice L can be compactly recorded in

- partially ordered set of join-irreducible elements $\langle \mathrm{Ji}(L), \leq \rangle$;
- the minimal join-covers of join-irreducible elements.

Pre-cursor of the *D*-basis

OD-graph of a finite lattice:
J.B.Nation, *An approach to lattice varieties of finite height*,
Alg. Universalis **27** (1990), 521–543.

The full information about finite lattice L can be compactly recorded in

- partially ordered set of join-irreducible elements $\langle \mathrm{Ji}(L), \leq \rangle$;
- the minimal join-covers of join-irreducible elements.

Example

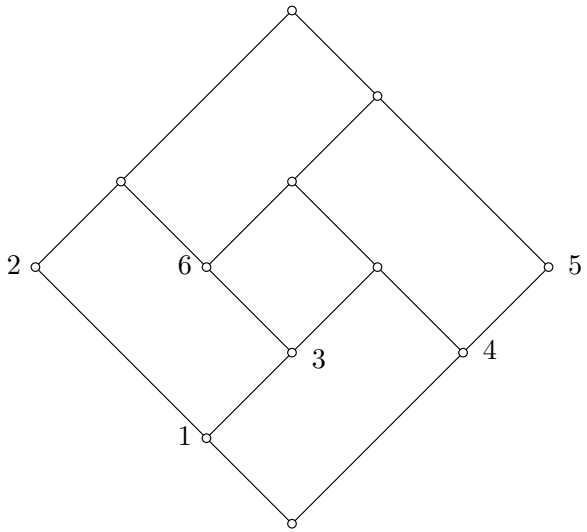


Figure: A_{12}

For lattice A_{12}, the poset of join-irreducible elements is:
$\langle \mathrm{Ji}(A_{12}), \leq \rangle = \langle \{1, 2, 3, 4, 5, 6, \}, 1 \leq 2, 1 \leq 3 \leq 6, 4 \leq 5 \rangle$.

Example continued

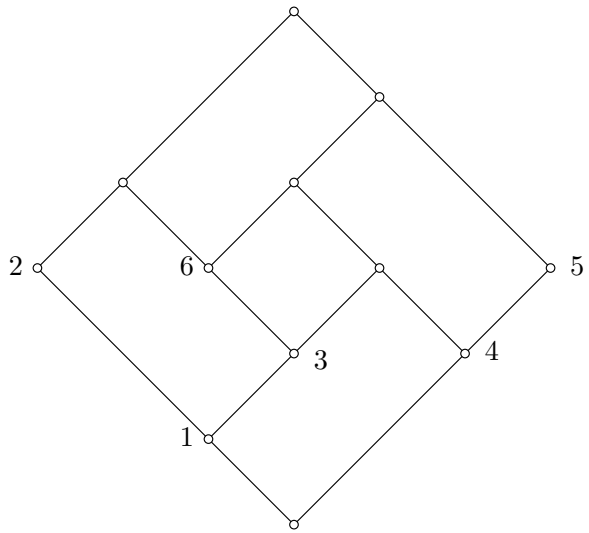
A *join-cover* is an expression $j \leq j_1 \vee \cdots \vee j_k$, $j \not\leq j_i$, $i \leq k$, for some $j, j_1, \ldots, j_k \in \mathrm{Ji}(L)$.

Examples: $3 \leq 1 \vee 4$, or $6 \leq 2 \vee 5$.

A join-cover $j \leq j_1 \vee \cdots \vee j_k$ is called *minimal*, if none of $j_1, \ldots, j_k$ can be replaced by smaller join-irreducibles or dropped so that one gets another join-cover.

$3 \leq 1 \vee 4$ is a minimal cover.

$6 \leq 2 \vee 5$ is not a minimal cover: since $4 \leq 5$ and $6 \leq 2 \vee 4$ is a cover.



Example continued

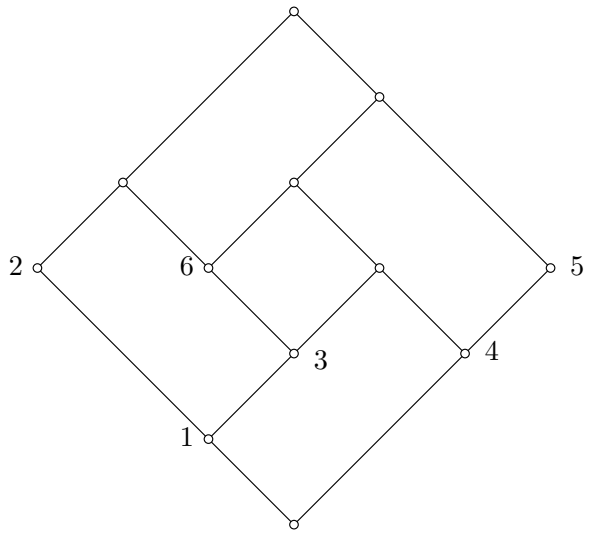
A *join-cover* is an expression $j \leq j_1 \vee \cdots \vee j_k$, $j \not\leq j_i$, $i \leq k$, for some $j, j_1, \ldots, j_k \in \mathrm{Ji}(L)$.

Examples: $3 \leq 1 \vee 4$, or $6 \leq 2 \vee 5$.

A join-cover $j \leq j_1 \vee \cdots \vee j_k$ is called *minimal*, if none of $j_1, \ldots, j_k$ can be replaced by smaller join-irreducibles or dropped so that one gets another join-cover.

$3 \leq 1 \vee 4$ is a minimal cover.

$6 \leq 2 \vee 5$ is not a minimal cover: since $4 \leq 5$ and $6 \leq 2 \vee 4$ is a cover.



Example continued

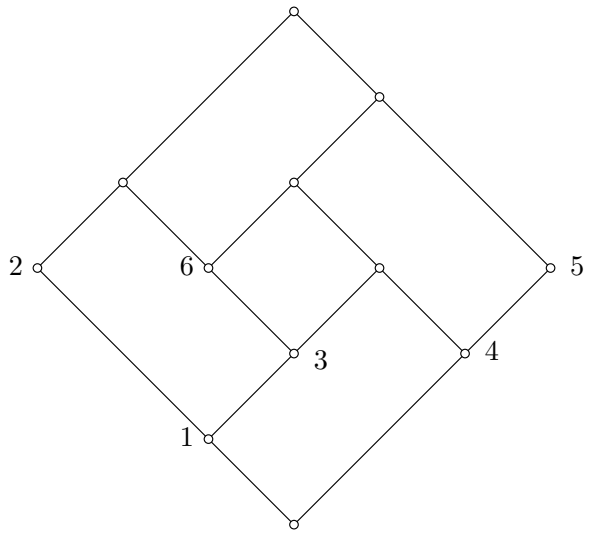
A *join-cover* is an expression $j \leq j_1 \vee \cdots \vee j_k$, $j \not\leq j_i$, $i \leq k$, for some $j, j_1, \ldots, j_k \in \mathrm{Ji}(L)$.

Examples: $3 \leq 1 \vee 4$, or $6 \leq 2 \vee 5$.

A join-cover $j \leq j_1 \vee \cdots \vee j_k$ is called *minimal*, if none of $j_1, \ldots, j_k$ can be replaced by smaller join-irreducibles or dropped so that one gets another join-cover.

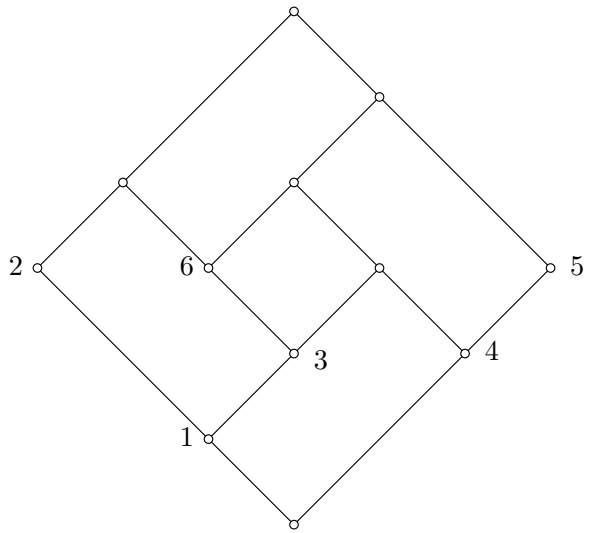
$3 \leq 1 \vee 4$ is a minimal cover.

$6 \leq 2 \vee 5$ is not a minimal cover: since $4 \leq 5$ and $6 \leq 2 \vee 4$ is a cover.



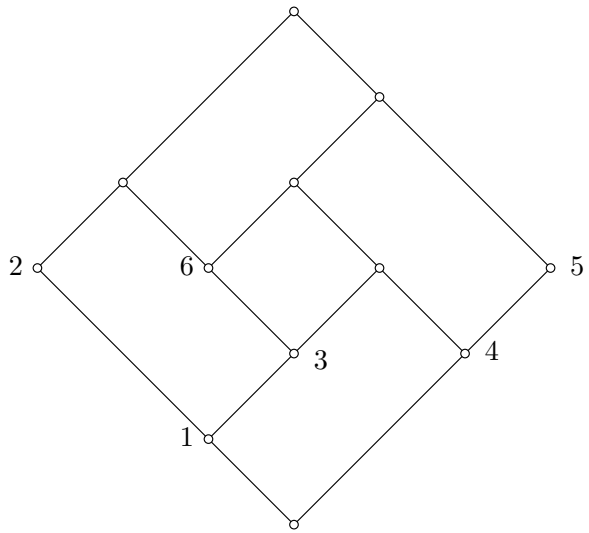
Example continued

A *join-cover* is an expression $j \leq j_1 \vee \cdots \vee j_k$, $j \not\leq j_i$, $i \leq k$, for some $j, j_1, \ldots, j_k \in \mathrm{Ji}(L)$.
Examples: $3 \leq 1 \vee 4$, or $6 \leq 2 \vee 5$.
A join-cover $j \leq j_1 \vee \cdots \vee j_k$ is called *minimal*, if none of $j_1, \ldots, j_k$ can be replaced by smaller join-irreducibles or dropped so that one gets another join-cover.
$3 \leq 1 \vee 4$ is a minimal cover.
$6 \leq 2 \vee 5$ is not a minimal cover: since $4 \leq 5$ and $6 \leq 2 \vee 4$ is a cover.



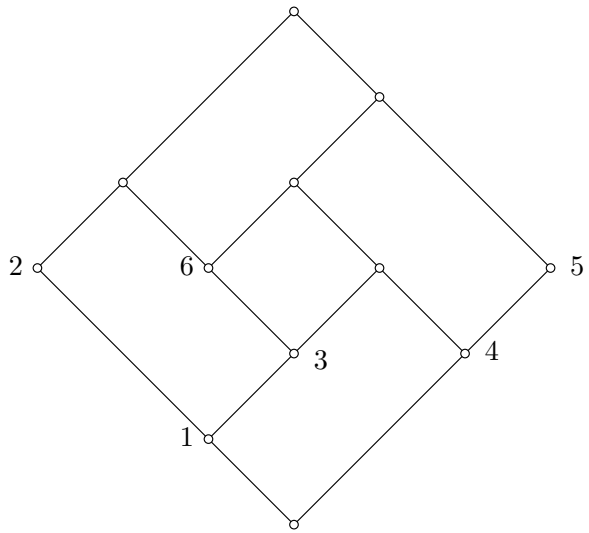
Example continued

A *join-cover* is an expression $j \leq j_1 \vee \cdots \vee j_k$, $j \not\leq j_i$, $i \leq k$, for some $j, j_1, \ldots, j_k \in \mathrm{Ji}(L)$.
Examples: $3 \leq 1 \vee 4$, or $6 \leq 2 \vee 5$.
A join-cover $j \leq j_1 \vee \cdots \vee j_k$ is called *minimal*, if none of $j_1, \ldots, j_k$ can be replaced by smaller join-irreducibles or dropped so that one gets another join-cover.
$3 \leq 1 \vee 4$ is a minimal cover.
$6 \leq 2 \vee 5$ is not a minimal cover: since $4 \leq 5$ and $6 \leq 2 \vee 4$ is a cover.



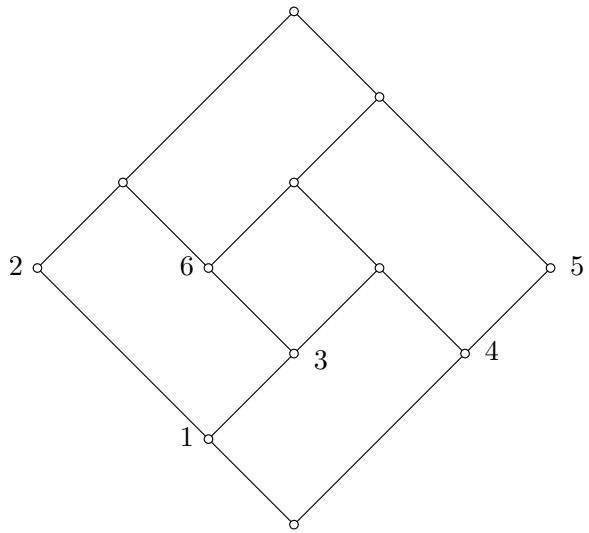
Example continued

A *join-cover* is an expression $j \leq j_1 \vee \cdots \vee j_k$, $j \not\leq j_i$, $i \leq k$, for some $j, j_1, \ldots, j_k \in \mathrm{Ji}(L)$.
Examples: $3 \leq 1 \vee 4$, or $6 \leq 2 \vee 5$.
A join-cover $j \leq j_1 \vee \cdots \vee j_k$ is called *minimal*, if none of $j_1, \ldots, j_k$ can be replaced by smaller join-irreducibles or dropped so that one gets another join-cover.
$3 \leq 1 \vee 4$ is a minimal cover.
$6 \leq 2 \vee 5$ is not a minimal cover: since $4 \leq 5$ and $6 \leq 2 \vee 4$ is a cover.



Example continued

A *join-cover* is an expression $j \leq j_1 \vee \cdots \vee j_k$, $j \not\leq j_i$, $i \leq k$, for some $j, j_1, \ldots, j_k \in \mathrm{Ji}(L)$.
Examples: $3 \leq 1 \vee 4$, or $6 \leq 2 \vee 5$.
A join-cover $j \leq j_1 \vee \cdots \vee j_k$ is called *minimal*, if none of $j_1, \ldots, j_k$ can be replaced by smaller join-irreducibles or dropped so that one gets another join-cover.
$3 \leq 1 \vee 4$ is a minimal cover.
$6 \leq 2 \vee 5$ is not a minimal cover: since $4 \leq 5$ and $6 \leq 2 \vee 4$ is a cover.



D-basis

The *D*-basis is introduced and studied in K. Adaricheva, J.B. Nation and R. Rand, *Ordered direct implicational basis of a finite closure system*, Disc. Appl. Math. **161** (2013), 707-723.

Definition

Let $\langle X, \phi \rangle$ be a standard closure system with $L = Cl(X, \phi)$.
The set of implications $\mathcal{S}_D$ is called the D-basis of $\langle X, \phi \rangle$, if it is made of two parts:

- $\{a \to b : b \in \phi(a)\}$*; equivalently, $b \leq a$ in $\langle Ji(L), \leq \rangle$.*
  This part is called a binary part of the basis.
- $\{j_1 \dots j_k \to j : j \leq j_1 \vee \dots \vee j_k$ *is a minimal cover in* $L\}$.

D-basis

The *D*-basis is introduced and studied in K. Adaricheva, J.B. Nation and R. Rand, *Ordered direct implicational basis of a finite closure system*, Disc. Appl. Math. **161** (2013), 707-723.

Definition

Let $\langle X, \phi \rangle$ be a standard closure system with $L = Cl(X, \phi)$.
The set of implications $\mathcal{S}_D$ is called the D-basis of $\langle X, \phi \rangle$, if it is made of two parts:

- $\{a \to b : b \in \phi(a)\}$*; equivalently, $b \leq a$ in $\langle Ji(L), \leq \rangle$.*
  This part is called a binary part of the basis.
- $\{j_1 \dots j_k \to j : j \leq j_1 \vee \dots \vee j_k$ *is a minimal cover in* $L\}$.

D-basis

The *D*-basis is introduced and studied in K. Adaricheva, J.B. Nation and R. Rand, *Ordered direct implicational basis of a finite closure system*, Disc. Appl. Math. **161** (2013), 707-723.

Definition

Let $\langle X, \phi \rangle$ be a standard closure system with $L = Cl(X, \phi)$.
The set of implications $\mathcal{S}_D$ is called the D-basis of $\langle X, \phi \rangle$, if it is made of two parts:

- $\{a \to b : b \in \phi(a)\}$*; equivalently, $b \leq a$ in $\langle Ji(L), \leq \rangle$.*
  This part is called a binary part of the basis.
- $\{j_1 \dots j_k \to j : j \leq j_1 \vee \dots \vee j_k$ *is a minimal cover in* $L\}$.

D-basis

The *D*-basis is introduced and studied in K. Adaricheva, J.B. Nation and R. Rand, *Ordered direct implicational basis of a finite closure system*, Disc. Appl. Math. **161** (2013), 707-723.

Definition

Let $\langle X, \phi \rangle$ be a standard closure system with $L = Cl(X, \phi)$.
The set of implications $\mathcal{S}_D$ is called the D-basis of $\langle X, \phi \rangle$, if it is made of two parts:

- $\{a \to b : b \in \phi(a)\}$*; equivalently, $b \leq a$ in $\langle Ji(L), \leq \rangle$.*
  This part is called a binary part of the basis.
- $\{j_1 \dots j_k \to j : j \leq j_1 \vee \dots \vee j_k$ *is a minimal cover in* $L\}$.

The *D*-relation and the *D*-basis

Why *D* in the name of the basis?

D-relation is an important concept in the study of free lattices, see R. Freese, J. Jezek, J.B. Nation “Free Lattices", 1995.

Definition

Given $b, c \in \mathrm{Ji}(L)$, one defines bDc, when there is a minimal cover $b \leq c \vee j_1 \vee \cdots \vee j_k$, for some $j_1, \ldots, j_k \in \mathrm{Ji}(L)$.

Equivalently: bDc iff there exists $Y \to b$ in the *D*-basis such that $c \in Y$.

Important: for every $Y \to b$ in the *D*-basis, $Y \subseteq bD = \{c \in \mathrm{Ji}(L) : bDc\}$.

The *D*-relation and the *D*-basis

Why *D* in the name of the basis?

D-relation is an important concept in the study of free lattices, see R. Freese, J. Jezek, J.B. Nation “Free Lattices", 1995.

Definition

Given $b, c \in \mathrm{Ji}(L)$, one defines bDc, when there is a minimal cover $b \leq c \vee j_1 \vee \cdots \vee j_k$, for some $j_1, \ldots, j_k \in \mathrm{Ji}(L)$.

Equivalently: bDc iff there exists $Y \to b$ in the *D*-basis such that $c \in Y$.

Important: for every $Y \to b$ in the *D*-basis, $Y \subseteq bD = \{c \in \mathrm{Ji}(L) : bDc\}$.

The *D*-relation and the *D*-basis

Why *D* in the name of the basis?

D-relation is an important concept in the study of free lattices, see R. Freese, J. Jezek, J.B. Nation “Free Lattices", 1995.

Definition

Given $b, c \in \mathrm{Ji}(L)$, one defines bDc, when there is a minimal cover $b \leq c \vee j_1 \vee \cdots \vee j_k$, for some $j_1, \ldots, j_k \in \mathrm{Ji}(L)$.

Equivalently: bDc iff there exists $Y \to b$ in the *D*-basis such that $c \in Y$.

Important: for every $Y \to b$ in the *D*-basis, $Y \subseteq bD = \{c \in \mathrm{Ji}(L) : bDc\}$.

The *D*-relation and the *D*-basis

Why *D* in the name of the basis?

D-relation is an important concept in the study of free lattices, see R. Freese, J. Jezek, J.B. Nation “Free Lattices", 1995.

Definition

Given $b, c \in \mathrm{Ji}(L)$*, one defines* bDc*, when there is a minimal cover* $b \leq c \vee j_1 \vee \cdots \vee j_k$*, for some* $j_1, \ldots, j_k \in \mathrm{Ji}(L)$*.*

Equivalently: bDc iff there exists $Y \to b$ in the *D*-basis such that $c \in Y$.

Important: for every $Y \to b$ in the *D*-basis, $Y \subseteq bD = \{c \in \mathrm{Ji}(L) : bDc\}$.

The *D*-relation and the *D*-basis

Why *D* in the name of the basis?

D-relation is an important concept in the study of free lattices, see R. Freese, J. Jezek, J.B. Nation “Free Lattices", 1995.

Definition

Given $b, c \in \mathrm{Ji}(L)$*, one defines* bDc*, when there is a minimal cover* $b \leq c \vee j_1 \vee \cdots \vee j_k$*, for some* $j_1, \ldots, j_k \in \mathrm{Ji}(L)$*.*

Equivalently: bDc iff there exists $Y \to b$ in the *D*-basis such that $c \in Y$.

Important: for every $Y \to b$ in the *D*-basis, $Y \subseteq bD = \{c \in \mathrm{Ji}(L) : bDc\}$.

The *D*-basis and the canonical unit basis

Theorem (ANR-2013)

- *S_D generates $\langle X, \phi \rangle$, i.e., D-basis is, indeed, a basis of this closure system.*
- *S_D is a subset of the canonical unit basis S_U.*

The *D*-basis and the canonical unit basis

Theorem (ANR-2013)

- *S_D generates $\langle X, \phi \rangle$, i.e., D-basis is, indeed, a basis of this closure system.*
- *S_D is a subset of the canonical unit basis S_U.*

The *D*-basis and the canonical unit basis

Theorem (ANR-2013)

- *S_D generates $\langle X, \phi \rangle$, i.e., D-basis is, indeed, a basis of this closure system.*
- *S_D is a subset of the canonical unit basis S_U.*

Comparison

Canonical direct unit basis $\mathcal{S}_U$ for $\langle \mathrm{Ji}(A_{12}), \phi \rangle$ has 13 implications.
$2 \to 1, 6 \to 1, 6 \to 3, 3 \to 1, 5 \to 4, 14 \to 3, 24 \to 3, 15 \to 3,$
$23 \to 6, 15 \to 6, 25 \to 6, 24 \to 5, 24 \to 6.$

D-basis has 9 implications.
$2 \to 1, 6 \to 3, 3 \to 1, 5 \to 4, 14 \to 3, 23 \to 6, 15 \to 6, 24 \to 5, 24 \to 6.$

Comparison

Canonical direct unit basis $\mathcal{S}_U$ for $\langle \mathrm{Ji}(A_{12}), \phi \rangle$ has 13 implications.
$2 \to 1,6 \to 1,6 \to 3,3 \to 1,5 \to 4,14 \to 3,24 \to 3,15 \to 3,$
$23 \to 6,15 \to 6,25 \to 6,24 \to 5,24 \to 6.$

D-basis has 9 implications.
$2 \to 1,6 \to 3,3 \to 1,5 \to 4,14 \to 3,23 \to 6,15 \to 6,24 \to 5,24 \to 6.$

Ordered iteration

Suppose the set of implications $\mathcal{S}$ are put into some linear order:

$$\mathcal{S} = \langle s_1, s_2, \ldots, s_n \rangle.$$

A mapping $\rho_{\mathcal{S}} : 2^X \to 2^X$ associated with this ordering is called an *ordered iteration* of $\mathcal{S}$:

- For any $Y \subseteq X$, let $Y_0 = Y$.
- If Y_k is computed and implication s_{k+1} is $A \to b$, then

$$Y_{k+1} = \begin{cases} Y_k \cup \{b\}, & \text{if } A \subseteq Y_k, \\ Y_k, & \text{otherwise.} \end{cases}$$

- Finally, $\rho_{\mathcal{S}}(Y) = Y_n$.

Definition

An implicational basis of $\langle X, \phi \rangle$, together with its order: $\mathcal{S} = \langle s_1, \ldots, s_n \rangle$ is called ordered direct, if $\rho(Y) = \phi(Y)$, for every $Y \subseteq X$.

Ordered iteration

Suppose the set of implications $\mathcal{S}$ are put into some linear order:

$$\mathcal{S} = \langle s_1, s_2, \ldots, s_n \rangle.$$

A mapping $\rho_{\mathcal{S}} : 2^X \to 2^X$ associated with this ordering is called an *ordered iteration* of $\mathcal{S}$:

- For any $Y \subseteq X$, let $Y_0 = Y$.
- If Y_k is computed and implication s_{k+1} is $A \to b$, then

$$Y_{k+1} = \begin{cases} Y_k \cup \{b\}, & \text{if } A \subseteq Y_k, \\ Y_k, & \text{otherwise.} \end{cases}$$

- Finally, $\rho_{\mathcal{S}}(Y) = Y_n$.

Definition

An implicational basis of $\langle X, \phi \rangle$, together with its order: $\mathcal{S} = \langle s_1, \ldots, s_n \rangle$ is called ordered direct, if $\rho(Y) = \phi(Y)$, for every $Y \subseteq X$.

Ordered iteration

Suppose the set of implications $\mathcal{S}$ are put into some linear order:

$$\mathcal{S} = \langle s_1, s_2, \ldots, s_n \rangle.$$

A mapping $\rho_{\mathcal{S}} : 2^X \to 2^X$ associated with this ordering is called an *ordered iteration* of $\mathcal{S}$:

- For any $Y \subseteq X$, let $Y_0 = Y$.
- If Y_k is computed and implication s_{k+1} is $A \to b$, then

$$Y_{k+1} = \begin{cases} Y_k \cup \{b\}, & \text{if } A \subseteq Y_k, \\ Y_k, & \text{otherwise.} \end{cases}$$

- Finally, $\rho_{\mathcal{S}}(Y) = Y_n$.

Definition

An implicational basis of $\langle X, \phi \rangle$, together with its order: $\mathcal{S} = \langle s_1, \ldots, s_n \rangle$ is called ordered direct, if $\rho(Y) = \phi(Y)$, for every $Y \subseteq X$.

Ordered iteration

Suppose the set of implications $\mathcal{S}$ are put into some linear order:

$$\mathcal{S} = \langle s_1, s_2, \ldots, s_n \rangle.$$

A mapping $\rho_{\mathcal{S}} : 2^X \to 2^X$ associated with this ordering is called an *ordered iteration* of $\mathcal{S}$:

- For any $Y \subseteq X$, let $Y_0 = Y$.
- If Y_k is computed and implication s_{k+1} is $A \to b$, then

$$Y_{k+1} = \begin{cases} Y_k \cup \{b\}, & \text{if } A \subseteq Y_k, \\ Y_k, & \text{otherwise.} \end{cases}$$

- Finally, $\rho_{\mathcal{S}}(Y) = Y_n$.

Definition

An implicational basis of $\langle X, \phi \rangle$, together with its order: $\mathcal{S} = \langle s_1, \ldots, s_n \rangle$ is called ordered direct, if $\rho(Y) = \phi(Y)$, for every $Y \subseteq X$.

Ordered iteration

Suppose the set of implications $\mathcal{S}$ are put into some linear order:

$$\mathcal{S} = \langle s_1, s_2, \ldots, s_n \rangle.$$

A mapping $\rho_{\mathcal{S}} : 2^X \to 2^X$ associated with this ordering is called an *ordered iteration* of $\mathcal{S}$:

- For any $Y \subseteq X$, let $Y_0 = Y$.
- If Y_k is computed and implication s_{k+1} is $A \to b$, then

$$Y_{k+1} = \begin{cases} Y_k \cup \{b\}, & \text{if } A \subseteq Y_k, \\ Y_k, & \text{otherwise.} \end{cases}$$

- Finally, $\rho_{\mathcal{S}}(Y) = Y_n$.

Definition

An implicational basis of $\langle X, \phi \rangle$, together with its order: $\mathcal{S} = \langle s_1, \ldots, s_n \rangle$ is called ordered direct, if $\rho(Y) = \phi(Y)$, for every $Y \subseteq X$.

Ordered iteration

Suppose the set of implications $\mathcal{S}$ are put into some linear order:

$$\mathcal{S} = \langle s_1, s_2, \ldots, s_n \rangle.$$

A mapping $\rho_{\mathcal{S}} : 2^X \to 2^X$ associated with this ordering is called an *ordered iteration* of $\mathcal{S}$:

- For any $Y \subseteq X$, let $Y_0 = Y$.
- If Y_k is computed and implication s_{k+1} is $A \to b$, then

$$Y_{k+1} = \begin{cases} Y_k \cup \{b\}, & \text{if } A \subseteq Y_k, \\ Y_k, & \text{otherwise.} \end{cases}$$

- Finally, $\rho_{\mathcal{S}}(Y) = Y_n$.

Definition

An implicational basis of $\langle X, \phi \rangle$, together with its order: $\mathcal{S} = \langle s_1, \ldots, s_n \rangle$ is called ordered direct, if $\rho(Y) = \phi(Y)$, for every $Y \subseteq X$.

Ordered direct basis

Theorem (ANR-2013)

- *$\mathcal{S}_D$ is an ordered direct basis, associated with any order, where the binary part precedes the rest of implications.*
- *There exist closure systems, for which the Guigues-Duquenne basis cannot be ordered.*

Algorithmic aspects

If $\mathcal{S}$ is a any unit direct basis of $\langle X, \phi \rangle$ of size $s = s(\mathcal{S})$ with m implications, then

- it takes time $O(s^2)$ to extract D-basis from $\mathcal{S}$;
- it takes time $O(m)$ to put extracted D-basis into a proper order.

Algorithmic aspects

If $\mathcal{S}$ is a any unit direct basis of $\langle X, \phi \rangle$ of size $s = s(\mathcal{S})$ with m implications, then

- it takes time $O(s^2)$ to extract D-basis from $\mathcal{S}$;
- it takes time $O(m)$ to put extracted D-basis into a proper order.

Algorithmic aspects

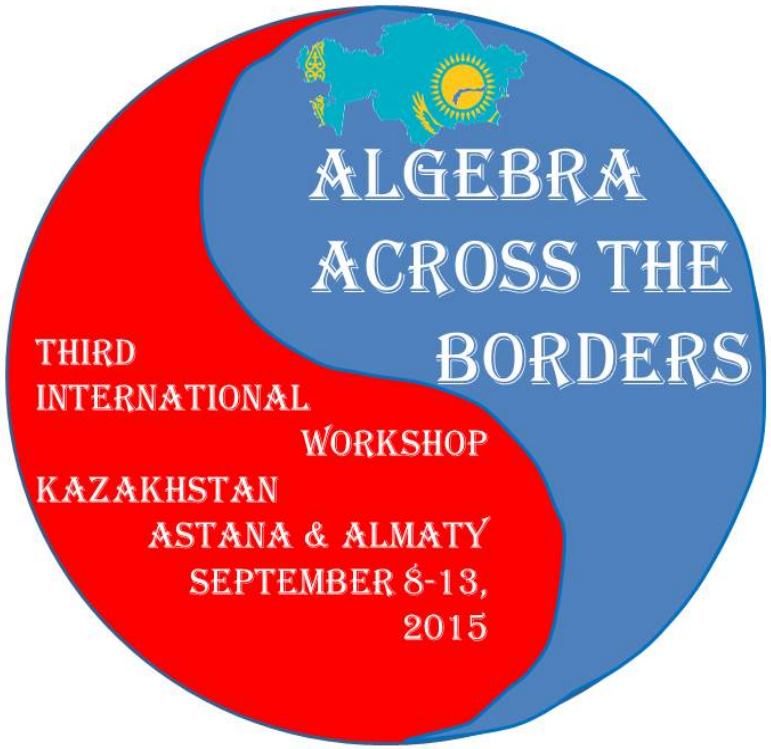
If $\mathcal{S}$ is a any unit direct basis of $\langle X, \phi \rangle$ of size $s = s(\mathcal{S})$ with m implications, then

- it takes time $O(s^2)$ to extract D-basis from $\mathcal{S}$;
- it takes time $O(m)$ to put extracted D-basis into a proper order.

Mid-talk conclusions

- For all practical purposes canonical direct unit basis can be replaced by the considerably shorter *D*-basis.
- The *D*-basis preserves the property of direct processing, assuming negligible pre-processing time for its ordering.

Algebra Across the Borders: workshop in Kazakhstan



Closure systems associated with a binary table

Table $\mathcal{T} = \langle U, A, R \rangle$, where $R \subseteq A \times U$ is a *relation* between U and A.

$S_A : 2^A \to 2^U$ is a *support function* on A

$S_A(Z) = \{y \in U : (z, y) \in R, \text{ for all } z \in Z\}$, for $Z \in 2^A$.

$S_U : 2^U \to 2^A$ is a *support function* on U

$S_U(Y) = \{z \in A : (z, y) \in R, \text{ for all } y \in Y\}$, for $Y \in 2^U$.

Lemma

Let $\mathcal{T} = \langle U, A, R \rangle$ be a table with support functions S_A and S_U.

- *S_A and S_U yield Galois connection between the power sets 2^A and 2^U.*
- *Mapping $\phi_A : Z \mapsto S_U(S_A(Z))$, with $Z \in 2^A$, is a closure operator on A.*
- *Similarly, mapping $\phi_U : Y \mapsto S_A(S_U(Y))$, with $Y \in 2^U$, is a closure operator on U.*

Closure systems associated with a binary table

Table $\mathcal{T} = \langle U, A, R \rangle$, where $R \subseteq A \times U$ is a *relation* between U and A.
$S_A : 2^A \to 2^U$ is a *support function* on A
$S_A(Z) = \{y \in U : (z, y) \in R, \text{ for all } z \in Z\}$, for $Z \in 2^A$.
$S_U : 2^U \to 2^A$ is a *support function* on U
$S_U(Y) = \{z \in A : (z, y) \in R, \text{ for all } y \in Y\}$, for $Y \in 2^U$.

Lemma

Let $\mathcal{T} = \langle U, A, R \rangle$ be a table with support functions S_A and S_U.

- *S_A and S_U yield Galois connection between the power sets 2^A and 2^U.*
- *Mapping $\phi_A : Z \mapsto S_U(S_A(Z))$, with $Z \in 2^A$, is a closure operator on A.*
- *Similarly, mapping $\phi_U : Y \mapsto S_A(S_U(Y))$, with $Y \in 2^U$, is a closure operator on U.*

Closure systems associated with a binary table

Table $\mathcal{T} = \langle U, A, R \rangle$, where $R \subseteq A \times U$ is a *relation* between U and A.
$S_A : 2^A \to 2^U$ is a *support function* on A
$S_A(Z) = \{y \in U : (z, y) \in R, \text{ for all } z \in Z\}$, for $Z \in 2^A$.
$S_U : 2^U \to 2^A$ is a *support function* on U
$S_U(Y) = \{z \in A : (z, y) \in R, \text{ for all } y \in Y\}$, for $Y \in 2^U$.

Lemma

Let $\mathcal{T} = \langle U, A, R \rangle$ be a table with support functions S_A and S_U.

- *S_A and S_U yield Galois connection between the power sets 2^A and 2^U.*
- *Mapping $\phi_A : Z \mapsto S_U(S_A(Z))$, with $Z \in 2^A$, is a closure operator on A.*
- *Similarly, mapping $\phi_U : Y \mapsto S_A(S_U(Y))$, with $Y \in 2^U$, is a closure operator on U.*

Closure systems associated with a binary table

Table $\mathcal{T} = \langle U, A, R \rangle$, where $R \subseteq A \times U$ is a *relation* between U and A.
$S_A : 2^A \to 2^U$ is a *support function* on A
$S_A(Z) = \{y \in U : (z, y) \in R, \text{ for all } z \in Z\}$, for $Z \in 2^A$.
$S_U : 2^U \to 2^A$ is a *support function* on U
$S_U(Y) = \{z \in A : (z, y) \in R, \text{ for all } y \in Y\}$, for $Y \in 2^U$.

Lemma

Let $\mathcal{T} = \langle U, A, R \rangle$ be a table with support functions S_A and S_U.

- *S_A and S_U yield Galois connection between the power sets 2^A and 2^U.*
- *Mapping $\phi_A : Z \mapsto S_U(S_A(Z))$, with $Z \in 2^A$, is a closure operator on A.*
- *Similarly, mapping $\phi_U : Y \mapsto S_A(S_U(Y))$, with $Y \in 2^U$, is a closure operator on U.*

Background

- G. Birkhoff, Lattice Theory, AMS Colloquium Publications **25** (1st ed), Providence, RI, 1940.
- M. Barbut and B. Monjardet, *Ordres et classifications: Algebre et combinatoire*, Hachette, Paris 1970.
- B. Ganter and R. Wille, *Formal Concept Analysis*, Mathematical foundations, Springer Ferlag, Berlin, 1999.

Background

- G. Birkhoff, Lattice Theory, AMS Colloquium Publications **25** (1st ed), Providence, RI, 1940.
- M. Barbut and B. Monjardet, *Ordres et classifications: Algebre et combinatoire*, Hachette, Paris 1970.
- B. Ganter and R. Wille, *Formal Concept Analysis*, Mathematical foundations, Springer Ferlag, Berlin, 1999.

Background

- G. Birkhoff, Lattice Theory, AMS Colloquium Publications **25** (1st ed), Providence, RI, 1940.
- M. Barbut and B. Monjardet, *Ordres et classifications: Algebre et combinatoire*, Hachette, Paris 1970.
- B. Ganter and R. Wille, *Formal Concept Analysis*, Mathematical foundations, Springer Ferlag, Berlin, 1999.

Implications and Association Rules in binary table

	C_1	C_2	DE	PDE	MP
1	1	1	0	0	1
2	1	0	1	1	0
3	1	0	1	1	0
4	0	1	0	1	0
5	0	1	1	0	1
6	0	1	0	1	1

- Implications: $(DE, PDE \to C_1)$, $(MP, DE \to C_2)$
- Association rules: $C_1 \to PDE$, with $s = 0.33$ and $c = 0.66$
  $C_2 \to MP$, with $s = 0.5$ and $c = 0.75$

Implications and Association Rules in binary table

	C_1	C_2	DE	PDE	MP
1	1	1	0	0	1
2	1	0	1	1	0
3	1	0	1	1	0
4	0	1	0	1	0
5	0	1	1	0	1
6	0	1	0	1	1

- Implications: $(DE, PDE \to C_1)$, $(MP, DE \to C_2)$
- Association rules: $C_1 \to PDE$, with $s = 0.33$ and $c = 0.66$
  $C_2 \to MP$, with $s = 0.5$ and $c = 0.75$

Implications and Association Rules in binary table

	C_1	C_2	DE	PDE	MP
1	1	1	0	0	1
2	1	0	1	1	0
3	1	0	1	1	0
4	0	1	0	1	0
5	0	1	1	0	1
6	0	1	0	1	1

- Implications: $(DE, PDE \rightarrow C_1)$, $(MP, DE \rightarrow C_2)$
- Association rules: $C_1 \rightarrow PDE$, with $s = 0.33$ and $c = 0.66$
  $C_2 \rightarrow MP$, with $s = 0.5$ and $c = 0.75$

Retrieval of association rules and implications from a binary table

To find $A \rightarrow B$ that holds in binary table

- as an implication in Guigues-Duquenne basis: identify A as a critical set for the closure system; left-side approach
- as an implication in D-basis: $B = \{b\}$, and one needs to find a minimal cover A for b; right-side approach
- as an association rule with $s > \alpha$, $c > \beta$: apply *Apriori* algorithm to find $A \cup B$ as a frequent set; homogeneous approach

Retrieval of association rules and implications from a binary table

To find $A \rightarrow B$ that holds in binary table

- as an implication in Guigues-Duquenne basis: identify A as a critical set for the closure system; left-side approach
- as an implication in D-basis: $B = \{b\}$, and one needs to find a minimal cover A for b; right-side approach
- as an association rule with $s > \alpha$, $c > \beta$: apply *Apriori* algorithm to find $A \cup B$ as a frequent set; homogeneous approach

Retrieval of association rules and implications from a binary table

To find $A \to B$ that holds in binary table

- as an implication in Guigues-Duquenne basis: identify A as a critical set for the closure system; left-side approach
- as an implication in D-basis: $B = \{b\}$, and one needs to find a minimal cover A for b; right-side approach
- as an association rule with $s > \alpha$, $c > \beta$: apply *Apriori* algorithm to find $A \cup B$ as a frequent set; homogeneous approach

Complexity

- Exponential delay in obtaining frequent sets: E. Boros, V. Gurvich, L. Khachiyan and K. Makino, *On the complexity of generating maximal frequent and minimal infrequent sets*, LNCS **2285** (2002), 133–141.
- For Guigues-Duquenne basis: all existing algorithms required generation of closure lattice, before attempting the basis retrieval.
- Series of results on (high) computational complexity of recognizing and enumerating critical sets: M. Babin and S. Kuznetsov (2010), F. Distel and B. Sertkaya (2011).

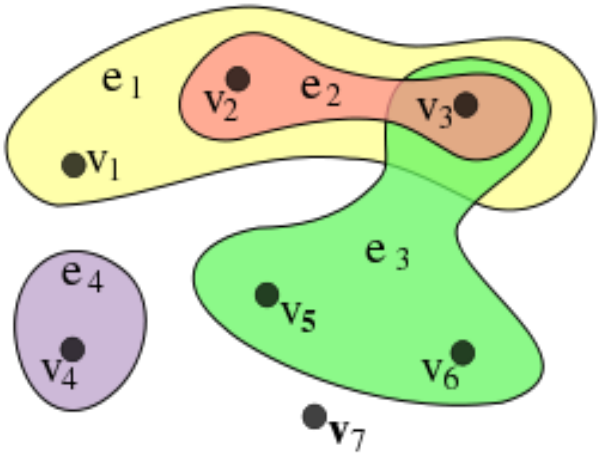
Retrieval of the canonical direct basis and the *D*-basis

- U. Ryssel, F. Distel and D. Borchmann, *Fast algorithms for implication bases and attribute exploration using proper premises*, Ann. Math. Art. Intell. **70** (2014), 25–53.
- K. Adaricheva, J.B. Nation, *Discovery of the D-basis in binary tables based on hypergraph dualization*, arxiv, subm. TCS, 2015.

Retrieval of the canonical direct basis and the *D*-basis

- U. Ryssel, F. Distel and D. Borchmann, *Fast algorithms for implication bases and attribute exploration using proper premises*, Ann. Math. Art. Intell. **70** (2014), 25–53.
- K. Adaricheva, J.B. Nation, *Discovery of the D-basis in binary tables based on hypergraph dualization*, arxiv, subm. TCS, 2015.

Hypergraph Dualization problem



- $V = \{v_1, \ldots, v_7\}$ is the set of vertices, $E = \{e_1, \ldots, e_4\} \subseteq \mathbf{2}^V$ is the set of hyper-edges.
- $H = \langle V, E \rangle$ is a hypergraph.
- $T \subseteq V$ is a *transversal*, if $T \cap e_i \neq \emptyset$, for all $e_i \in E$.
- Problem: find all *minimal* transversals of given hypergraph H.
- Solution: $H^d = \{V, E^d = \{v_4v_3, v_4v_2v_5, v_4v_2v_6\}\}$ is a dual hypergraph.

Algorithmic solutions to Hypergraph Dualization

- M. Fredman and L. Khachiyan, *On the complexity of dualization of monotone disjunctive forms*, J. Algorithms **21** (1996), 618–628. Problem of generating all minimal transversals can be solved in time $O(N^{o(\log N)})$ time, where N is the size of input and output.
- Test results of code implementation of algorithm are presented in L. Khachiyan, E. Boros, K. Elbassioni and V. Gurvich, Disc. Appl. Math. **154** (2006), 2350–2372.
- Recent implementation: K. Murakami and T. Uno, *Efficient algorithms for generating large scale hypergraphs*, Disc. Appl. Math. **170** (2014), 83–94.

Algorithmic solutions to Hypergraph Dualization

- M. Fredman and L. Khachiyan, *On the complexity of dualization of monotone disjunctive forms*, J. Algorithms **21** (1996), 618–628. Problem of generating all minimal transversals can be solved in time $O(N^{o(\log N)})$ time, where N is the size of input and output.
- Test results of code implementation of algorithm are presented in L. Khachiyan, E. Boros, K. Elbassioni and V. Gurvich, Disc. Appl. Math. **154** (2006), 2350–2372.
- Recent implementation: K. Murakami and T. Uno, *Efficient algorithms for generating large scale hypergraphs*, Disc. Appl. Math. **170** (2014), 83–94.

Algorithmic solutions to Hypergraph Dualization

- M. Fredman and L. Khachiyan, *On the complexity of dualization of monotone disjunctive forms*, J. Algorithms **21** (1996), 618–628. Problem of generating all minimal transversals can be solved in time $O(N^{o(\log N)})$ time, where N is the size of input and output.
- Test results of code implementation of algorithm are presented in L. Khachiyan, E. Boros, K. Elbassioni and V. Gurvich, Disc. Appl. Math. **154** (2006), 2350–2372.
- Recent implementation: K. Murakami and T. Uno, *Efficient algorithms for generating large scale hypergraphs*, Disc. Appl. Math. **170** (2014), 83–94.

Algorithmic solutions to Hypergraph Dualization

- M. Fredman and L. Khachiyan, *On the complexity of dualization of monotone disjunctive forms*, J. Algorithms **21** (1996), 618–628. Problem of generating all minimal transversals can be solved in time $O(N^{o(\log N)})$ time, where N is the size of input and output.
- Test results of code implementation of algorithm are presented in L. Khachiyan, E. Boros, K. Elbassioni and V. Gurvich, Disc. Appl. Math. **154** (2006), 2350–2372.
- Recent implementation: K. Murakami and T. Uno, *Efficient algorithms for generating large scale hypergraphs*, Disc. Appl. Math. **170** (2014), 83–94.

Motivation for the right-side approach

	C_1	C_2	*DE*	*PDE*	*MP*
1	1	1	0	0	1
2	1	0	1	1	0
3	1	0	1	1	0
4	0	1	0	1	0
5	0	1	1	0	1
6	0	1	0	1	1

Implication: $MP, PDE \rightarrow C_2$

Motivation for the right-side approach

	C_1	C_2	*DE*	*PDE*	*MP*
1	1	1	0	0	1
2	1	0	1	1	0
3	1	0	1	1	0
4	0	1	0	1	0
5	0	1	1	0	1
6	0	1	0	1	1

Implication: $MP, PDE \rightarrow C_2$

Instance of HD problem for the D-basis retrieval

- Fix $b \in A$, one particular attribute. The goal: obtain all $Y \to b$ from the D-basis.
- Due to the definition of the D-basis, all such Y are subsets of $bD = \{c \in A : bDc\}$.
- Use Lemma 11.10 from *Free Lattices* book: bDc, for $b, c \in \mathrm{Ji}(L)$ iff there exists $p \in \mathrm{Mi}(L)$ such that $b \uparrow p$ and $p \downarrow c$.
- Attributes of the table play the role of join-irreducibles and the objects the role of meet-irreducibles of the concept lattice.
- $\uparrow$ and $\downarrow$ relations between the attributes and objects of the table can be found in time polynomial in the size of the table.
- Hypergraph associated with the fixed $b \in A$: set of vertices $V = bD$; hyperedges are $H_p = \{c \in bD : cRp\}$, for each $p \in U$, for which $b \uparrow p$.

Instance of HD problem for the D-basis retrieval

- Fix $b \in A$, one particular attribute. The goal: obtain all $Y \to b$ from the D-basis.
- Due to the definition of the D-basis, all such Y are subsets of $bD = \{c \in A : bDc\}$.
- Use Lemma 11.10 from *Free Lattices* book: bDc, for $b, c \in \mathrm{Ji}(L)$ iff there exists $p \in \mathrm{Mi}(L)$ such that $b \uparrow p$ and $p \downarrow c$.
- Attributes of the table play the role of join-irreducibles and the objects the role of meet-irreducibles of the concept lattice.
- $\uparrow$ and $\downarrow$ relations between the attributes and objects of the table can be found in time polynomial in the size of the table.
- Hypergraph associated with the fixed $b \in A$: set of vertices $V = bD$; hyperedges are $H_p = \{c \in bD : cRp\}$, for each $p \in U$, for which $b \uparrow p$.

Instance of HD problem for the D-basis retrieval

- Fix $b \in A$, one particular attribute. The goal: obtain all $Y \to b$ from the D-basis.
- Due to the definition of the D-basis, all such Y are subsets of $bD = \{c \in A : bDc\}$.
- Use Lemma 11.10 from *Free Lattices* book: bDc, for $b, c \in \mathrm{Ji}(L)$ iff there exists $p \in \mathrm{Mi}(L)$ such that $b \uparrow p$ and $p \downarrow c$.
- Attributes of the table play the role of join-irreducibles and the objects the role of meet-irreducibles of the concept lattice.
- $\uparrow$ and $\downarrow$ relations between the attributes and objects of the table can be found in time polynomial in the size of the table.
- Hypergraph associated with the fixed $b \in A$: set of vertices $V = bD$; hyperedges are $H_p = \{c \in bD : cRp\}$, for each $p \in U$, for which $b \uparrow p$.

Instance of HD problem for the D-basis retrieval

- Fix $b \in A$, one particular attribute. The goal: obtain all $Y \to b$ from the D-basis.
- Due to the definition of the D-basis, all such Y are subsets of $bD = \{c \in A : bDc\}$.
- Use Lemma 11.10 from *Free Lattices* book: bDc, for $b, c \in \mathrm{Ji}(L)$ iff there exists $p \in \mathrm{Mi}(L)$ such that $b \uparrow p$ and $p \downarrow c$.
- Attributes of the table play the role of join-irreducibles and the objects the role of meet-irreducibles of the concept lattice.
- $\uparrow$ and $\downarrow$ relations between the attributes and objects of the table can be found in time polynomial in the size of the table.
- Hypergraph associated with the fixed $b \in A$: set of vertices $V = bD$; hyperedges are $H_p = \{c \in bD : cRp\}$, for each $p \in U$, for which $b \uparrow p$.

Instance of HD problem for the D-basis retrieval

- Fix $b \in A$, one particular attribute. The goal: obtain all $Y \to b$ from the D-basis.
- Due to the definition of the D-basis, all such Y are subsets of $bD = \{c \in A : bDc\}$.
- Use Lemma 11.10 from *Free Lattices* book: bDc, for $b, c \in \mathrm{Ji}(L)$ iff there exists $p \in \mathrm{Mi}(L)$ such that $b \uparrow p$ and $p \downarrow c$.
- Attributes of the table play the role of join-irreducibles and the objects the role of meet-irreducibles of the concept lattice.
- $\uparrow$ and $\downarrow$ relations between the attributes and objects of the table can be found in time polynomial in the size of the table.
- Hypergraph associated with the fixed $b \in A$: set of vertices $V = bD$; hyperedges are $H_p = \{c \in bD : cRp\}$, for each $p \in U$, for which $b \uparrow p$.

Instance of HD problem for the D-basis retrieval

- Fix $b \in A$, one particular attribute. The goal: obtain all $Y \to b$ from the D-basis.
- Due to the definition of the D-basis, all such Y are subsets of $bD = \{c \in A : bDc\}$.
- Use Lemma 11.10 from *Free Lattices* book: bDc, for $b, c \in \mathrm{Ji}(L)$ iff there exists $p \in \mathrm{Mi}(L)$ such that $b \uparrow p$ and $p \downarrow c$.
- Attributes of the table play the role of join-irreducibles and the objects the role of meet-irreducibles of the concept lattice.
- $\uparrow$ and $\downarrow$ relations between the attributes and objects of the table can be found in time polynomial in the size of the table.
- Hypergraph associated with the fixed $b \in A$: set of vertices $V = bD$; hyperedges are $H_p = \{c \in bD : cRp\}$, for each $p \in U$, for which $b \uparrow p$.

Astana-New York-Honolulu-Tokyo Project

- In May, 2013: the first working code implementation of *D*-basis retrieval from the binary table via HD, written by students J. Blumenkopf and T. Moldwin (Yeshiva University, New York). It implemented the call to existing code of T. Uno (Tokyo).
- In June, 2013: retrieved the *D*-basis of about 49,000 implications from 50-by-100 matrix of density 0.2, in 3 min and 30 sec.
- In March 2015 the it took 49 hours to retrieve more than 2,000,000 implications of the *D*-basis pertinent to one attribute in 61-by-287 matrix of density 0.35, with the medical data from Cancer research lab in Astana.
- One needs to work further with 2,000,000 implications to make sense out of it.
- This work is related to sorting the *association rules* in data mining (José Luiz Balcázar in his talk on Thursday!)

Astana-New York-Honolulu-Tokyo Project

- In May, 2013: the first working code implementation of *D*-basis retrieval from the binary table via HD, written by students J. Blumenkopf and T. Moldwin (Yeshiva University, New York). It implemented the call to existing code of T. Uno (Tokyo).
- In June, 2013: retrieved the *D*-basis of about 49,000 implications from 50-by-100 matrix of density 0.2, in 3 min and 30 sec.
- In March 2015 the it took 49 hours to retrieve more than 2,000,000 implications of the *D*-basis pertinent to one attribute in 61-by-287 matrix of density 0.35, with the medical data from Cancer research lab in Astana.
- One needs to work further with 2,000,000 implications to make sense out of it.
- This work is related to sorting the *association rules* in data mining (José Luiz Balcázar in his talk on Thursday!)

Astana-New York-Honolulu-Tokyo Project

- In May, 2013: the first working code implementation of *D*-basis retrieval from the binary table via HD, written by students J. Blumenkopf and T. Moldwin (Yeshiva University, New York). It implemented the call to existing code of T. Uno (Tokyo).
- In June, 2013: retrieved the *D*-basis of about 49,000 implications from 50-by-100 matrix of density 0.2, in 3 min and 30 sec.
- In March 2015 the it took 49 hours to retrieve more than 2,000,000 implications of the *D*-basis pertinent to one attribute in 61-by-287 matrix of density 0.35, with the medical data from Cancer research lab in Astana.
- One needs to work further with 2,000,000 implications to make sense out of it.
- This work is related to sorting the *association rules* in data mining (José Luiz Balcázar in his talk on Thursday!)

Astana-New York-Honolulu-Tokyo Project

- In May, 2013: the first working code implementation of *D*-basis retrieval from the binary table via HD, written by students J. Blumenkopf and T. Moldwin (Yeshiva University, New York). It implemented the call to existing code of T. Uno (Tokyo).
- In June, 2013: retrieved the *D*-basis of about 49,000 implications from 50-by-100 matrix of density 0.2, in 3 min and 30 sec.
- In March 2015 the it took 49 hours to retrieve more than 2,000,000 implications of the *D*-basis pertinent to one attribute in 61-by-287 matrix of density 0.35, with the medical data from Cancer research lab in Astana.
- One needs to work further with 2,000,000 implications to make sense out of it.
- This work is related to sorting the *association rules* in data mining (José Luiz Balcázar in his talk on Thursday!)

Astana-New York-Honolulu-Tokyo Project

- In May, 2013: the first working code implementation of *D*-basis retrieval from the binary table via HD, written by students J. Blumenkopf and T. Moldwin (Yeshiva University, New York). It implemented the call to existing code of T. Uno (Tokyo).
- In June, 2013: retrieved the *D*-basis of about 49,000 implications from 50-by-100 matrix of density 0.2, in 3 min and 30 sec.
- In March 2015 the it took 49 hours to retrieve more than 2,000,000 implications of the *D*-basis pertinent to one attribute in 61-by-287 matrix of density 0.35, with the medical data from Cancer research lab in Astana.
- One needs to work further with 2,000,000 implications to make sense out of it.
- This work is related to sorting the *association rules* in data mining (José Luiz Balcázar in his talk on Thursday!)

Filtering of association rules

Main challenge: how to separate the significant rules from less interesting?

Parameters for filtering are versions and combinations of the *support* and the *confidence*.

Implications are association rules of confidence = 1.

Filtering of association rules

Main challenge: how to separate the significant rules from less interesting?
Parameters for filtering are versions and combinations of the *support* and the *confidence*.
Implications are association rules of confidence = 1.

Filtering of association rules

Main challenge: how to separate the significant rules from less interesting?
Parameters for filtering are versions and combinations of the *support* and the *confidence*.
Implications are association rules of confidence = 1.

New measurement of implications: the relevance

Suppose attribute b and all implications $X \to b$ from the D-basis are of particular interest. For any other attribute x we may compute *total support of* x in relation to b:

$$\text{tsup}_b(x) = Sum\{\frac{\sup(X \to b)}{|X|} : x \in X\}$$

The behavior of the same attribute a in relation to negation of attribute b can be also computed, if one replaces column b by its complement $\neg b$:

$$\text{tsup}_{\neg b}(x) = Sum\{\frac{\sup(X \to \neg b)}{|X|} : x \in X\}$$

Finally, the *relevance* of attribute a in relation to b will be the ratio of two total supports above:

$$\text{rel}_b(x) = \frac{\text{tsup}_b(x)}{\text{tsup}_{\neg b}(x) + 1}.$$

New measurement of implications: the relevance

Suppose attribute b and all implications $X \to b$ from the D-basis are of particular interest. For any other attribute x we may compute *total support of* x in relation to b:

$$\text{tsup}_b(x) = Sum\{\frac{\sup(X \to b)}{|X|} : x \in X\}$$

The behavior of the same attribute a in relation to negation of attribute b can be also computed, if one replaces column b by its complement $\neg b$:

$$\text{tsup}_{\neg b}(x) = Sum\{\frac{\sup(X \to \neg b)}{|X|} : x \in X\}$$

Finally, the *relevance* of attribute a in relation to b will be the ratio of two total supports above:

$$\text{rel}_b(x) = \frac{\text{tsup}_b(x)}{\text{tsup}_{\neg b}(x) + 1}.$$

New measurement of implications: the relevance

Suppose attribute b and all implications $X \to b$ from the D-basis are of particular interest. For any other attribute x we may compute *total support of* x in relation to b:

$$\text{tsup}_b(x) = Sum\{\frac{\sup(X \to b)}{|X|} : x \in X\}$$

The behavior of the same attribute a in relation to negation of attribute b can be also computed, if one replaces column b by its complement $\neg b$:

$$\text{tsup}_{\neg b}(x) = Sum\{\frac{\sup(X \to \neg b)}{|X|} : x \in X\}$$

Finally, the *relevance* of attribute a in relation to b will be the ratio of two total supports above:

$$\text{rel}_b(x) = \frac{\text{tsup}_b(x)}{\text{tsup}_{\neg b}(x) + 1}.$$

Ranking the attributes and implications related to attribute b

- Relevance allows to rank all attributes of the table, in relation to one fixed attribute b;
- the larger the value of the relevance, the more important attribute x for presence of attribute b; only values $\mathrm{rel}_b(x) > 1$ should be of interest;
- one can rank implications $X \to b$ by computing the average value of $\mathrm{rel}_b(x)$, for $x \in X$; the implications with the largest average value of the relevance could be of most importance.

Ranking the attributes and implications related to attribute b

- Relevance allows to rank all attributes of the table, in relation to one fixed attribute b;
- the larger the value of the relevance, the more important attribute x for presence of attribute b; only values $\mathrm{rel}_b(x) > 1$ should be of interest;
- one can rank implications $X \to b$ by computing the average value of $\mathrm{rel}_b(x)$, for $x \in X$; the implications with the largest average value of the relevance could be of most importance.

Ranking the attributes and implications related to attribute b

- Relevance allows to rank all attributes of the table, in relation to one fixed attribute b;
- the larger the value of the relevance, the more important attribute x for presence of attribute b; only values $\mathrm{rel}_b(x) > 1$ should be of interest;
- one can rank implications $X \to b$ by computing the average value of $\mathrm{rel}_b(x)$, for $x \in X$; the implications with the largest average value of the relevance could be of most importance.

Medical data testing

- two data sets from medical studies;
- in Honolulu: collaboration with bio-informatics groups of Cancer center, which provided gene expression data of 291 ovarian cancer patients;
- in Astana: collaboration with medical group applying new regimen of treatment to the group of 61 brain cancer patients.

Medical data testing

- two data sets from medical studies;
- in Honolulu: collaboration with bio-informatics groups of Cancer center, which provided gene expression data of 291 ovarian cancer patients;
- in Astana: collaboration with medical group applying new regimen of treatment to the group of 61 brain cancer patients.

Medical data testing

- two data sets from medical studies;
- in Honolulu: collaboration with bio-informatics groups of Cancer center, which provided gene expression data of 291 ovarian cancer patients;
- in Astana: collaboration with medical group applying new regimen of treatment to the group of 61 brain cancer patients.

Gene expression in ovarian cancer data set

- Data set is composed of genes found to be highly correlated with microRNA and DNA methylation in a common set of 291 ovarian tumor samples;
- total number of genes in original data set was over 16,000;
- global gene expression, microRNA and DNA methylation data for each tumor sample were downloaded from The Cancer Genome Atlas (TCGA);
- each sample also had meta-data composed of censored time-to-death from all causes;
- the resulting data tables were jointly analyzed using matrix factorizations to identify the dominant source of variation in the data as a sparse linear model;
- hierarchical clustering and pathway analysis methods were then employed to identify an even smaller set of genes that models the dominant signal as a sparse linear combination.

Gene expression in ovarian cancer data set

- Data set is composed of genes found to be highly correlated with microRNA and DNA methylation in a common set of 291 ovarian tumor samples;
- total number of genes in original data set was over 16,000;
- global gene expression, microRNA and DNA methylation data for each tumor sample were downloaded from The Cancer Genome Atlas (TCGA);
- each sample also had meta-data composed of censored time-to-death from all causes;
- the resulting data tables were jointly analyzed using matrix factorizations to identify the dominant source of variation in the data as a sparse linear model;
- hierarchical clustering and pathway analysis methods were then employed to identify an even smaller set of genes that models the dominant signal as a sparse linear combination.

Gene expression in ovarian cancer data set

- Data set is composed of genes found to be highly correlated with microRNA and DNA methylation in a common set of 291 ovarian tumor samples;
- total number of genes in original data set was over 16,000;
- global gene expression, microRNA and DNA methylation data for each tumor sample were downloaded from The Cancer Genome Atlas (TCGA);
- each sample also had meta-data composed of censored time-to-death from all causes;
- the resulting data tables were jointly analyzed using matrix factorizations to identify the dominant source of variation in the data as a sparse linear model;
- hierarchical clustering and pathway analysis methods were then employed to identify an even smaller set of genes that models the dominant signal as a sparse linear combination.

Gene expression in ovarian cancer data set

- Data set is composed of genes found to be highly correlated with microRNA and DNA methylation in a common set of 291 ovarian tumor samples;
- total number of genes in original data set was over 16,000;
- global gene expression, microRNA and DNA methylation data for each tumor sample were downloaded from The Cancer Genome Atlas (TCGA);
- each sample also had meta-data composed of censored time-to-death from all causes;
- the resulting data tables were jointly analyzed using matrix factorizations to identify the dominant source of variation in the data as a sparse linear model;
- hierarchical clustering and pathway analysis methods were then employed to identify an even smaller set of genes that models the dominant signal as a sparse linear combination.

Gene expression in ovarian cancer data set

- Data set is composed of genes found to be highly correlated with microRNA and DNA methylation in a common set of 291 ovarian tumor samples;
- total number of genes in original data set was over 16,000;
- global gene expression, microRNA and DNA methylation data for each tumor sample were downloaded from The Cancer Genome Atlas (TCGA);
- each sample also had meta-data composed of censored time-to-death from all causes;
- the resulting data tables were jointly analyzed using matrix factorizations to identify the dominant source of variation in the data as a sparse linear model;
- hierarchical clustering and pathway analysis methods were then employed to identify an even smaller set of genes that models the dominant signal as a sparse linear combination.

Relevance computation for subsets of genes

- Two important ”signature" subsets were identified by combination of methods: one with 21 genes and another with 40 genes;
- real-valued data was converted to the binary one using two columns for each gene: one for over-expressed values, and another for under-expressed
- survival data was coded in 4 columns: two were representing the upper half and lower half, and two others the upper and lower quartiles of the entire group of 291 patients

Relevance computation for subsets of genes

- Two important ”signature" subsets were identified by combination of methods: one with 21 genes and another with 40 genes;
- real-valued data was converted to the binary one using two columns for each gene: one for over-expressed values, and another for under-expressed
- survival data was coded in 4 columns: two were representing the upper half and lower half, and two others the upper and lower quartiles of the entire group of 291 patients

Relevance computation for subsets of genes

- Two important ”signature" subsets were identified by combination of methods: one with 21 genes and another with 40 genes;
- real-valued data was converted to the binary one using two columns for each gene: one for over-expressed values, and another for under-expressed
- survival data was coded in 4 columns: two were representing the upper half and lower half, and two others the upper and lower quartiles of the entire group of 291 patients

6-gene signature of ovarian cancer

- For the set of 21 genes (targets of OSM), the relevance computation was done for the columns of upper half and lower half of survival attributes;
- the six genes with the highest relevance to long survival (over 1300 days) turned out to be IL1B, VDR, SELE, HLA-B, GBP2 and IL15RA;
- the difference between the Kaplan-Meier plots for the top and bottom quartiles of the 291 training samples ordered by the 6-gene *D*-basis signature turned out statistically significant in both the Kaplan-Meier analysis (p=0.00217) and Cox regression analysis (p=0.0000269);
- the same analysis on the 99 validation samples gave p=0.0176 for the KM analysis and p=0.0419 for the Cox regression analysis.

6-gene signature of ovarian cancer

- For the set of 21 genes (targets of OSM), the relevance computation was done for the columns of upper half and lower half of survival attributes;
- the six genes with the highest relevance to long survival (over 1300 days) turned out to be IL1B, VDR, SELE, HLA-B, GBP2 and IL15RA;
- the difference between the Kaplan-Meier plots for the top and bottom quartiles of the 291 training samples ordered by the 6-gene *D*-basis signature turned out statistically significant in both the Kaplan-Meier analysis (p=0.00217) and Cox regression analysis (p=0.0000269);
- the same analysis on the 99 validation samples gave p=0.0176 for the KM analysis and p=0.0419 for the Cox regression analysis.

6-gene signature of ovarian cancer

- For the set of 21 genes (targets of OSM), the relevance computation was done for the columns of upper half and lower half of survival attributes;
- the six genes with the highest relevance to long survival (over 1300 days) turned out to be IL1B, VDR, SELE, HLA-B, GBP2 and IL15RA;
- the difference between the Kaplan-Meier plots for the top and bottom quartiles of the 291 training samples ordered by the 6-gene *D*-basis signature turned out statistically significant in both the Kaplan-Meier analysis (p=0.00217) and Cox regression analysis (p=0.0000269);
- the same analysis on the 99 validation samples gave p=0.0176 for the KM analysis and p=0.0419 for the Cox regression analysis.

6-gene signature of ovarian cancer

- For the set of 21 genes (targets of OSM), the relevance computation was done for the columns of upper half and lower half of survival attributes;
- the six genes with the highest relevance to long survival (over 1300 days) turned out to be IL1B, VDR, SELE, HLA-B, GBP2 and IL15RA;
- the difference between the Kaplan-Meier plots for the top and bottom quartiles of the 291 training samples ordered by the 6-gene *D*-basis signature turned out statistically significant in both the Kaplan-Meier analysis (p=0.00217) and Cox regression analysis (p=0.0000269);
- the same analysis on the 99 validation samples gave p=0.0176 for the KM analysis and p=0.0419 for the Cox regression analysis.

6-gene signature of ovarian cancer

- On the set of 40 genes *several* relevance computations was performed; for long survival: for the column with top quartile and column with top half; also, *modified relevance* of top quartile versus lowest quartile;
- similar computations were done for the low survival columns;
- the genes with the *combination* of high relevance scores in all three computations were identified, one group for the high survival and another for the low survival;
- the six genes (from this set of 40 targets of IL4) with the highest relevance to long survival turned out to be FCGR2A, CD86, IFI30, CCL5, SELPLG and ICOS;
- the difference between the KM plots for the top and bottom quartiles on both: the training and 99 validation samples data - ordered by the 6-gene D-basis signature turned out to be statistically significant in both the KM analysis (p=0.0143, p=0.0137) and Cox regression analysis (p=0.00112,p=0.00858).

6-gene signature of ovarian cancer

- On the set of 40 genes *several* relevance computations was performed; for long survival: for the column with top quartile and column with top half; also, *modified relevance* of top quartile versus lowest quartile;
- similar computations were done for the low survival columns;
- the genes with the *combination* of high relevance scores in all three computations were identified, one group for the high survival and another for the low survival;
- the six genes (from this set of 40 targets of IL4) with the highest relevance to long survival turned out to be FCGR2A, CD86, IFI30, CCL5, SELPLG and ICOS;
- the difference between the KM plots for the top and bottom quartiles on both: the training and 99 validation samples data - ordered by the 6-gene D-basis signature turned out to be statistically significant in both the KM analysis (p=0.0143, p=0.0137) and Cox regression analysis (p=0.00112,p=0.00858).

6-gene signature of ovarian cancer

- On the set of 40 genes *several* relevance computations was performed; for long survival: for the column with top quartile and column with top half; also, *modified relevance* of top quartile versus lowest quartile;
- similar computations were done for the low survival columns;
- the genes with the *combination* of high relevance scores in all three computations were identified, one group for the high survival and another for the low survival;
- the six genes (from this set of 40 targets of IL4) with the highest relevance to long survival turned out to be FCGR2A, CD86, IFI30, CCL5, SELPLG and ICOS;
- the difference between the KM plots for the top and bottom quartiles on both: the training and 99 validation samples data - ordered by the 6-gene D-basis signature turned out to be statistically significant in both the KM analysis (p=0.0143, p=0.0137) and Cox regression analysis (p=0.00112,p=0.00858).

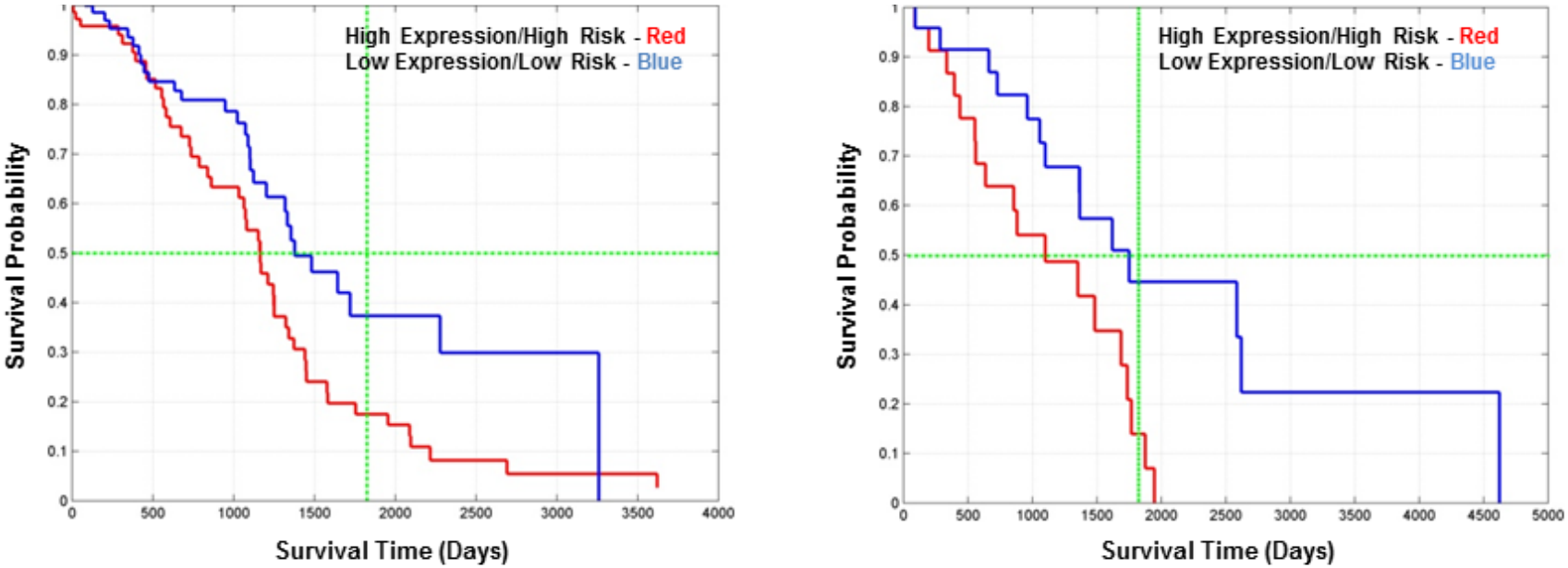
6-gene signature of ovarian cancer

- On the set of 40 genes *several* relevance computations was performed; for long survival: for the column with top quartile and column with top half; also, *modified relevance* of top quartile versus lowest quartile;
- similar computations were done for the low survival columns;
- the genes with the *combination* of high relevance scores in all three computations were identified, one group for the high survival and another for the low survival;
- the six genes (from this set of 40 targets of IL4) with the highest relevance to long survival turned out to be FCGR2A, CD86, IFI30, CCL5, SELPLG and ICOS;
- the difference between the KM plots for the top and bottom quartiles on both: the training and 99 validation samples data - ordered by the 6-gene D-basis signature turned out to be statistically significant in both the KM analysis (p=0.0143, p=0.0137) and Cox regression analysis (p=0.00112,p=0.00858).

6-gene signature of ovarian cancer

- On the set of 40 genes *several* relevance computations was performed; for long survival: for the column with top quartile and column with top half; also, *modified relevance* of top quartile versus lowest quartile;
- similar computations were done for the low survival columns;
- the genes with the *combination* of high relevance scores in all three computations were identified, one group for the high survival and another for the low survival;
- the six genes (from this set of 40 targets of IL4) with the highest relevance to long survival turned out to be FCGR2A, CD86, IFI30, CCL5, SELPLG and ICOS;
- the difference between the KM plots for the top and bottom quartiles on both: the training and 99 validation samples data - ordered by the 6-gene D-basis signature turned out to be statistically significant in both the KM analysis (p=0.0143, p=0.0137) and Cox regression analysis (p=0.00112,p=0.00858).

Kaplan-Meier Plots



Kaplan-Meier (KM) plots of training and independent test data stratified by the 6-gene D-base signature (DBSig6). Panel A. KM plots of top and bottom quartiles of 291 training samples ordered by DBSig6. Note that DBSig6 was extracted from a subset of the 291 training samples used to derive the KM plots. The Blue KM plot models the survival of the lowest quartile of patients ordered by DBSig6 expression and red plot models patients in highest quartile. The difference in KM plots is statistically significant (logrankP = 0.0143) and DBSig6 expression is associated with survival even after adjustment for age and stage (CoxP = 0.0011). The intersection of the horizontal green line with the two KM curves gives the median survival time for each group. The intersection of the vertical green line with the KM curves gives the 5-year survival rate for each group. **Panel B**. The Interpretation of the blue and red KM plots and the horizontal and vertical green lines are the same as in Panel A. The difference in KM plots on the test data is statistically significant (logrankP = 0.0137) and DBSig6 expression is predictive of survival after adjustment for age and stage (CoxP = 0.0086). This result demonstrates that the DBSig6 signature is able to generalize to data unseen during discovery.

Astana medical data analysis

- The original set of data for 61 patients with brain tumors (astrocytomas, glioblastomas, and meningiomas) under new regimen of treatment was collected in the hospital of Medical Holding in Astana, between 2012–2013.
- Three groups of parameters were regularly measured in patients. *The first group* was set of flow cytometry markers to identify major immune cell populations in peripheral blood;
- *the second group* was blood analysis for creatinine, bilirubin, calcium, protein, amylase etc.;
- *the third group* of parameters was infectious markers including indicators for hepatitis A and C, cytomegalovirus, chlamydia, herpes simplex, etc.
- *the main read-out* measurement of patient response to the new treatment was clinical assessment: column C1 coded the group of patients succumbing to illness, column C2 coded the patients with stabilizing or improving condition.

Astana medical data analysis

- The original set of data for 61 patients with brain tumors (astrocytomas, glioblastomas, and meningiomas) under new regimen of treatment was collected in the hospital of Medical Holding in Astana, between 2012–2013.
- Three groups of parameters were regularly measured in patients. *The first group* was set of flow cytometry markers to identify major immune cell populations in peripheral blood;
- *the second group* was blood analysis for creatinine, bilirubin, calcium, protein, amylase etc.;
- *the third group* of parameters was infectious markers including indicators for hepatitis A and C, cytomegalovirus, chlamydia, herpes simplex, etc.
- *the main read-out* measurement of patient response to the new treatment was clinical assessment: column C1 coded the group of patients succumbing to illness, column C2 coded the patients with stabilizing or improving condition.

Astana medical data analysis

- The original set of data for 61 patients with brain tumors (astrocytomas, glioblastomas, and meningiomas) under new regimen of treatment was collected in the hospital of Medical Holding in Astana, between 2012–2013.
- Three groups of parameters were regularly measured in patients. *The first group* was set of flow cytometry markers to identify major immune cell populations in peripheral blood;
- *the second group* was blood analysis for creatinine, bilirubin, calcium, protein, amylase etc.;
- *the third group* of parameters was infectious markers including indicators for hepatitis A and C, cytomegalovirus, chlamydia, herpes simplex, etc.
- *the main read-out* measurement of patient response to the new treatment was clinical assessment: column C1 coded the group of patients succumbing to illness, column C2 coded the patients with stabilizing or improving condition.

Astana medical data analysis

- The original set of data for 61 patients with brain tumors (astrocytomas, glioblastomas, and meningiomas) under new regimen of treatment was collected in the hospital of Medical Holding in Astana, between 2012–2013.
- Three groups of parameters were regularly measured in patients. *The first group* was set of flow cytometry markers to identify major immune cell populations in peripheral blood;
- *the second group* was blood analysis for creatinine, bilirubin, calcium, protein, amylase etc.;
- *the third group* of parameters was infectious markers including indicators for hepatitis A and C, cytomegalovirus, chlamydia, herpes simplex, etc.
- *the main read-out* measurement of patient response to the new treatment was clinical assessment: column C1 coded the group of patients succumbing to illness, column C2 coded the patients with stabilizing or improving condition.

Astana medical data analysis

- The original set of data for 61 patients with brain tumors (astrocytomas, glioblastomas, and meningiomas) under new regimen of treatment was collected in the hospital of Medical Holding in Astana, between 2012–2013.
- Three groups of parameters were regularly measured in patients. *The first group* was set of flow cytometry markers to identify major immune cell populations in peripheral blood;
- *the second group* was blood analysis for creatinine, bilirubin, calcium, protein, amylase etc.;
- *the third group* of parameters was infectious markers including indicators for hepatitis A and C, cytomegalovirus, chlamydia, herpes simplex, etc.
- *the main read-out* measurement of patient response to the new treatment was clinical assessment: column C1 coded the group of patients succumbing to illness, column C2 coded the patients with stabilizing or improving condition.

Conversion to binary data

- Temporal data for most parameters: several measurements during period of treatment;
- various length of treatment;
- for consistency of analysis, the data was converted to incremental form, using three points of period of treatment;
- two columns were used to code the increments: one for *increasing* pattern, another for *decreasing* pattern;
- some initial values were used, converting to 4 columns of quartiles;
- resulting table had 61 rows and 278 columns;
- more testing was done of a subgroup of 33 patients with specific type of cancer.

Conversion to binary data

- Temporal data for most parameters: several measurements during period of treatment;
- various length of treatment;
- for consistency of analysis, the data was converted to incremental form, using three points of period of treatment;
- two columns were used to code the increments: one for *increasing* pattern, another for *decreasing* pattern;
- some initial values were used, converting to 4 columns of quartiles;
- resulting table had 61 rows and 278 columns;
- more testing was done of a subgroup of 33 patients with specific type of cancer.

Conversion to binary data

- Temporal data for most parameters: several measurements during period of treatment;
- various length of treatment;
- for consistency of analysis, the data was converted to incremental form, using three points of period of treatment;
- two columns were used to code the increments: one for *increasing* pattern, another for *decreasing* pattern;
- some initial values were used, converting to 4 columns of quartiles;
- resulting table had 61 rows and 278 columns;
- more testing was done of a subgroup of 33 patients with specific type of cancer.

Conversion to binary data

- Temporal data for most parameters: several measurements during period of treatment;
- various length of treatment;
- for consistency of analysis, the data was converted to incremental form, using three points of period of treatment;
- two columns were used to code the increments: one for *increasing* pattern, another for *decreasing* pattern;
- some initial values were used, converting to 4 columns of quartiles;
- resulting table had 61 rows and 278 columns;
- more testing was done of a subgroup of 33 patients with specific type of cancer.

Conversion to binary data

- Temporal data for most parameters: several measurements during period of treatment;
- various length of treatment;
- for consistency of analysis, the data was converted to incremental form, using three points of period of treatment;
- two columns were used to code the increments: one for *increasing* pattern, another for *decreasing* pattern;
- some initial values were used, converting to 4 columns of quartiles;
- resulting table had 61 rows and 278 columns;
- more testing was done of a subgroup of 33 patients with specific type of cancer.

Conversion to binary data

- Temporal data for most parameters: several measurements during period of treatment;
- various length of treatment;
- for consistency of analysis, the data was converted to incremental form, using three points of period of treatment;
- two columns were used to code the increments: one for *increasing* pattern, another for *decreasing* pattern;
- some initial values were used, converting to 4 columns of quartiles;
- resulting table had 61 rows and 278 columns;
- more testing was done of a subgroup of 33 patients with specific type of cancer.

Conversion to binary data

- Temporal data for most parameters: several measurements during period of treatment;
- various length of treatment;
- for consistency of analysis, the data was converted to incremental form, using three points of period of treatment;
- two columns were used to code the increments: one for *increasing* pattern, another for *decreasing* pattern;
- some initial values were used, converting to 4 columns of quartiles;
- resulting table had 61 rows and 278 columns;
- more testing was done of a subgroup of 33 patients with specific type of cancer.

Relevance computation

- The request for computation of the basis for column (C2) resulted in 1,138,518 implications computed in 39639 sec, or just over 11 hours;
- for the column indicating the group of declining patients (C1) the number of implications was 2,073,282, and it was computed in 170458 sec, or 47.34 hours;
- much higher variation of the relevance parameter in the case of 61 patients, than in test for 33 patients;
- most of the highly relevant attributes in the test on the subgroup of 33 patients showed their significance in the test for 61 as well;
- on the set of 33 patients, 3 highly ranked attributes for column C1 were the dynamics of the same immune parameter: CD3+CD8+ cytotoxic T cells during two halves of treatment and the whole period of treatment;
- in the C2 group, highly relevant were attributes for decreasing dynamics of the presence of two specific viruses.

Relevance computation

- The request for computation of the basis for column (C2) resulted in 1,138,518 implications computed in 39639 sec, or just over 11 hours;
- for the column indicating the group of declining patients (C1) the number of implications was 2,073,282, and it was computed in 170458 sec, or 47.34 hours;
- much higher variation of the relevance parameter in the case of 61 patients, than in test for 33 patients;
- most of the highly relevant attributes in the test on the subgroup of 33 patients showed their significance in the test for 61 as well;
- on the set of 33 patients, 3 highly ranked attributes for column C1 were the dynamics of the same immune parameter: CD3+CD8+ cytotoxic T cells during two halves of treatment and the whole period of treatment;
- in the C2 group, highly relevant were attributes for decreasing dynamics of the presence of two specific viruses.

Relevance computation

- The request for computation of the basis for column (C2) resulted in 1,138,518 implications computed in 39639 sec, or just over 11 hours;
- for the column indicating the group of declining patients (C1) the number of implications was 2,073,282, and it was computed in 170458 sec, or 47.34 hours;
- much higher variation of the relevance parameter in the case of 61 patients, than in test for 33 patients;
- most of the highly relevant attributes in the test on the subgroup of 33 patients showed their significance in the test for 61 as well;
- on the set of 33 patients, 3 highly ranked attributes for column C1 were the dynamics of the same immune parameter: CD3+CD8+ cytotoxic T cells during two halves of treatment and the whole period of treatment;
- in the C2 group, highly relevant were attributes for decreasing dynamics of the presence of two specific viruses.

Relevance computation

- The request for computation of the basis for column (C2) resulted in 1,138,518 implications computed in 39639 sec, or just over 11 hours;
- for the column indicating the group of declining patients (C1) the number of implications was 2,073,282, and it was computed in 170458 sec, or 47.34 hours;
- much higher variation of the relevance parameter in the case of 61 patients, than in test for 33 patients;
- most of the highly relevant attributes in the test on the subgroup of 33 patients showed their significance in the test for 61 as well;
- on the set of 33 patients, 3 highly ranked attributes for column C1 were the dynamics of the same immune parameter: CD3+CD8+ cytotoxic T cells during two halves of treatment and the whole period of treatment;
- in the C2 group, highly relevant were attributes for decreasing dynamics of the presence of two specific viruses.

Relevance computation

- The request for computation of the basis for column (C2) resulted in 1,138,518 implications computed in 39639 sec, or just over 11 hours;
- for the column indicating the group of declining patients (C1) the number of implications was 2,073,282, and it was computed in 170458 sec, or 47.34 hours;
- much higher variation of the relevance parameter in the case of 61 patients, than in test for 33 patients;
- most of the highly relevant attributes in the test on the subgroup of 33 patients showed their significance in the test for 61 as well;
- on the set of 33 patients, 3 highly ranked attributes for column C1 were the dynamics of the same immune parameter: CD3+CD8+ cytotoxic T cells during two halves of treatment and the whole period of treatment;
- in the C2 group, highly relevant were attributes for decreasing dynamics of the presence of two specific viruses.

Relevance computation

- The request for computation of the basis for column (C2) resulted in 1,138,518 implications computed in 39639 sec, or just over 11 hours;
- for the column indicating the group of declining patients (C1) the number of implications was 2,073,282, and it was computed in 170458 sec, or 47.34 hours;
- much higher variation of the relevance parameter in the case of 61 patients, than in test for 33 patients;
- most of the highly relevant attributes in the test on the subgroup of 33 patients showed their significance in the test for 61 as well;
- on the set of 33 patients, 3 highly ranked attributes for column C1 were the dynamics of the same immune parameter: CD3+CD8+ cytotoxic T cells during two halves of treatment and the whole period of treatment;
- in the C2 group, highly relevant were attributes for decreasing dynamics of the presence of two specific viruses.

Statistical analysis of real-valued data for 61 patients

- Statistical analysis of the immune, biochemical and infection parameters: initial values, averages and increments - was also performed using *non-parametric* Spearman's correlation analysis;
- for analysis of the parameters' difference between groups of patients, the *non-parametric* Kruskal-Wallis test in the R language for statistical computing were applied.

Statistical analysis of real-valued data for 61 patients

- Statistical analysis of the immune, biochemical and infection parameters: initial values, averages and increments - was also performed using *non-parametric* Spearman's correlation analysis;
- for analysis of the parameters' difference between groups of patients, the *non-parametric* Kruskal-Wallis test in the R language for statistical computing were applied.

Parameters associated with patients' clinical assessment

Table 2. Parameters associated with patients' clinical assessment.

Parameters *	Group 1	Group 2	Group 3	Group 4	P-value
INF_V1_HBsAg	0.540	0.573	0.572	0.427	0.015
INF_Avg_HBsAg	0.576	0.496	0.519	0.335	0.025
IMM_V1_CD3+CD8+	36.470	33.100	30.740	26.220	0.025
BLD_V1_IgG	9.910	7.100	12.660	9.070	0.007
BLD_V1_Fe_serum	8.290	16.570	13.205	24.030	0.030
BLD_δ31_triglycerids	0.175	-0.277	-0.222	-0.250	0.022
BLD_δ31_HDL	-0.330	0.052	0.030	0.011	0.002
BLD_δ31_LDL	-0.572	-0.230	-0.419	-0.116	0.011
BLD_δ31_Creatinin	-0.213	0.132	0.217	0.240	0.006
BLD_δ31_Total_protein	-0.109	0.012	0.016	-0.033	0.030
BLD_δ31_Albumin	-0.257	0.017	0.026	-0.130	0.035
BLD_δ31_CRP	4.399	0.337	-0.496	0.140	0.039
BLD_δ31_IgA	0.348	0.239	-0.092	-0.232	0.011
BLD_δ31_Lipase	-0.594	-0.294	0.127	nd	0.019
BLD_δ31_Fe_serum	-0.608	-0.437	0.251	-0.540	0.020
IMM_δ31_CD3-CD19+	-0.500	-0.300	0.000	-0.056	0.027
IMM_δ32_CD3+CD4+	-0.363	-0.106	0.076	0.176	0.013

* **BLD** – patients' blood parameters, **IMM** – immune parameters, **INF** – infection parameters. **HBsAg** - hepatitis B virus surface antigen, **CD3+CD8+** - cytotoxic T cells levels in peripheral blood, **IgG** – total immunoglobulins, **Fe_serum** - iron in blood, **HDL** – high density lipoproteins, **LDL** – low density lipoproteins, **CRP** - C-reactive protein, **IgA** – immunoglobulin of IgA isotype, **CD3-CD19+** - level of B cells in peripheral blood, **CD3+CD4+** - level of T helper cells. **Group** –median

Parameters common for correlation and high relevance with survival

Table 3. Parameters discovered with both implications and statistical approaches.

Parameters *	relevance	group
BLD_δ31_Total_protein_inc10	4.12	C1
BLD_δ31_HDL_inc10	2.45	C1
BLD_δ31_CRP_inc10	2.33	C1
IMM_δ32_CD3+CD4+_inc10	2.12	C1
IMM_δ32_CD3-CD19+_inc10	1.87	C1
BLD_ δ31_Total_protein_dec10	1.85	C1
BLD_δ31_CRP_dec10	1.81	C1
BLD_δ31_LDL_dec10	1.62	C1
INF_Avg_HBsAg_dec10	1.59	C1
BLD_δ31_triglycerids_dec10	1.59	C1
BLD_δ31_Creatinin_inc10	1.53	C1
IMM_δ32_CD3-CD19+_dec10	1.52	C1
BLD_δ31_Albumin_dec10	6.87	C2
BLD_δ31_Fe_serum_inc10	4.43	C2
BLD_δ31_IgA_inc10	2.15	C2
BLD_δ31_Creatinin_dec10	1.91	C2
BLD_δ31_triglycerids_inc10	1.81	C2

Venn diagram of high relevance parameters and parameters correlated with survival

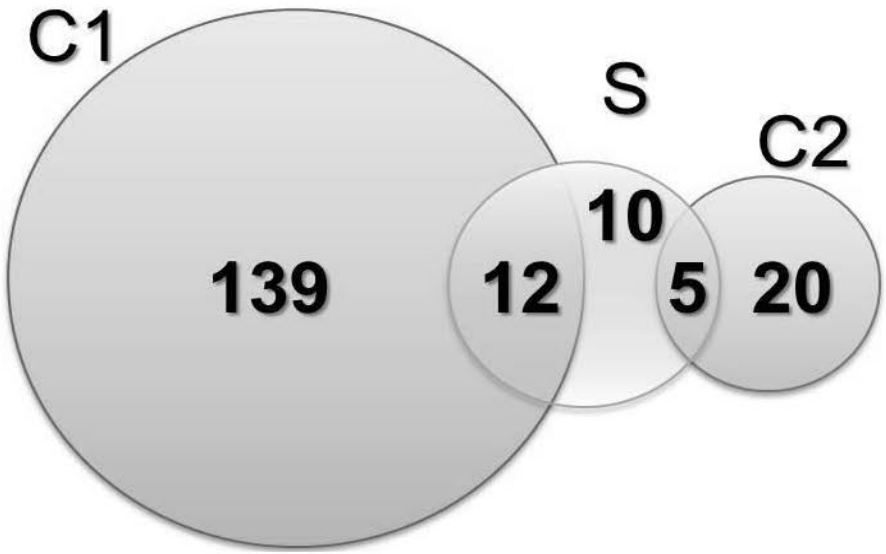


Figure 1. Venn diagram for overlapping results of implication and statistical approaches. Circle **C1** - group of declining patients with total of 139+12=151 implications. Circle **C2** – group of improving patients with total of 20+5=25 implications. Threshold for relevance is set at 1.5. Circle **S** – results of statistical analysis with total of 12+10+5= 27 significant biases at $p < 0.05$ threshold. See Tables 3 and 4 for details.

Further analysis and research

- Further testing will be done on additional validation sample obtained through the observation of patients in 2014;
- more computation of the relevance will be done on the columns different from survival;
- new tests will be performed with the data on other types of cancer.

Further analysis and research

- Further testing will be done on additional validation sample obtained through the observation of patients in 2014;
- more computation of the relevance will be done on the columns different from survival;
- new tests will be performed with the data on other types of cancer.

Further analysis and research

- Further testing will be done on additional validation sample obtained through the observation of patients in 2014;
- more computation of the relevance will be done on the columns different from survival;
- new tests will be performed with the data on other types of cancer.

Future work

For the discovery of all association rules with high threshold of the confidence c:

- run algorithm of D-basis retrieval on a batch of matrices obtained by removing $100(1-c)$ rows from original data set;
- take the union of all obtained implications: they will be the association rules of the confidence at least c;
- one can retrieve only rules $Y \to b$, for a fixed attribute of interest b.

Conclusions

- new algorithm of retrieval of basis of implications proves to be most time-effective compared to other existing algorithms;
- computation of implications meets the same challenge as Data Mining field: staggering amount of output
- to manage the output and rank the attributes with respect to a fixed attribute b, a new measure of the *relevance* can be computed;
- it allows the choice of the most important implications which could be provided to specialists on the data for further analysis;
- application of the method: two new 6-gene signatures for ovarian cancer were selected within targeted subsets of 21 and 40 genes;
- new method can be complementary to existing techniques, in analysis of complex diseases such as cancer.

Conclusions

- new algorithm of retrieval of basis of implications proves to be most time-effective compared to other existing algorithms;
- computation of implications meets the same challenge as Data Mining field: staggering amount of output
- to manage the output and rank the attributes with respect to a fixed attribute b, a new measure of the *relevance* can be computed;
- it allows the choice of the most important implications which could be provided to specialists on the data for further analysis;
- application of the method: two new 6-gene signatures for ovarian cancer were selected within targeted subsets of 21 and 40 genes;
- new method can be complementary to existing techniques, in analysis of complex diseases such as cancer.

Conclusions

- new algorithm of retrieval of basis of implications proves to be most time-effective compared to other existing algorithms;
- computation of implications meets the same challenge as Data Mining field: staggering amount of output
- to manage the output and rank the attributes with respect to a fixed attribute *b*, a new measure of the *relevance* can be computed;
- it allows the choice of the most important implications which could be provided to specialists on the data for further analysis;
- application of the method: two new 6-gene signatures for ovarian cancer were selected within targeted subsets of 21 and 40 genes;
- new method can be complementary to existing techniques, in analysis of complex diseases such as cancer.

Conclusions

- new algorithm of retrieval of basis of implications proves to be most time-effective compared to other existing algorithms;
- computation of implications meets the same challenge as Data Mining field: staggering amount of output
- to manage the output and rank the attributes with respect to a fixed attribute *b*, a new measure of the *relevance* can be computed;
- it allows the choice of the most important implications which could be provided to specialists on the data for further analysis;
- application of the method: two new 6-gene signatures for ovarian cancer were selected within targeted subsets of 21 and 40 genes;
- new method can be complementary to existing techniques, in analysis of complex diseases such as cancer.

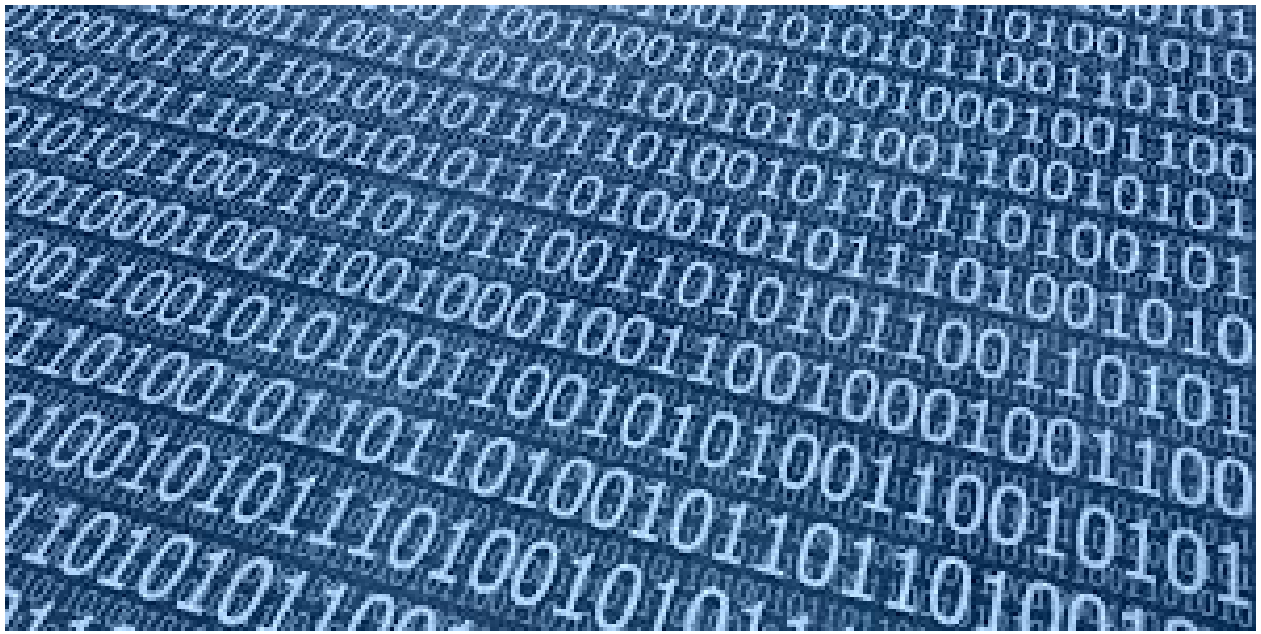
Conclusions

- new algorithm of retrieval of basis of implications proves to be most time-effective compared to other existing algorithms;
- computation of implications meets the same challenge as Data Mining field: staggering amount of output
- to manage the output and rank the attributes with respect to a fixed attribute b, a new measure of the *relevance* can be computed;
- it allows the choice of the most important implications which could be provided to specialists on the data for further analysis;
- application of the method: two new 6-gene signatures for ovarian cancer were selected within targeted subsets of 21 and 40 genes;
- new method can be complementary to existing techniques, in analysis of complex diseases such as cancer.

Conclusions

- new algorithm of retrieval of basis of implications proves to be most time-effective compared to other existing algorithms;
- computation of implications meets the same challenge as Data Mining field: staggering amount of output
- to manage the output and rank the attributes with respect to a fixed attribute b, a new measure of the *relevance* can be computed;
- it allows the choice of the most important implications which could be provided to specialists on the data for further analysis;
- application of the method: two new 6-gene signatures for ovarian cancer were selected within targeted subsets of 21 and 40 genes;
- new method can be complementary to existing techniques, in analysis of complex diseases such as cancer.

Thank you for your attendance and attention!



3d International Workshop "Algebra across the borders"

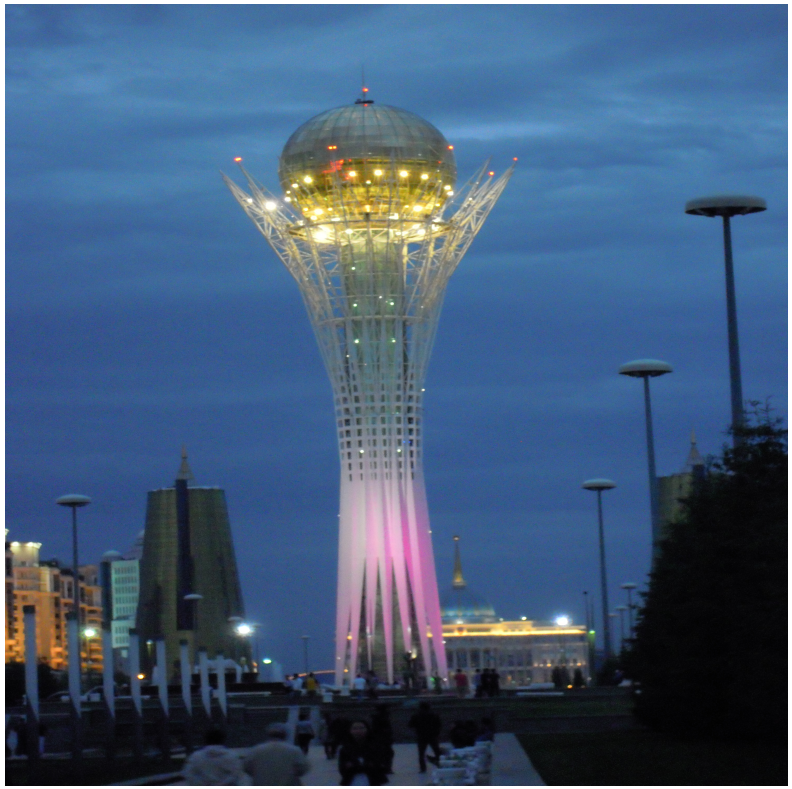
The workshop program will start on September 8-10, 2015, Tuesday to Thursday, in Nazarbayev University, in Astana, the new capital of Kazakhstan.

*Pictures: courtesy of J.B Nation

3d International Workshop "Algebra across the borders"

The workshop program will start on September 8-10, 2015, Tuesday to Thursday, in Nazarbayev University, in Astana, the new capital of Kazakhstan.

*Pictures: courtesy of J.B Nation



3d International Workshop "Algebra across the borders"

The workshop program will start on September 8-10, 2015, Tuesday to Thursday, in Nazarbayev University, in Astana, the new capital of Kazakhstan.

*Pictures: courtesy of J.B Nation

3d International Workshop "Algebra across the borders"

Then we relocate to the Almaty region, for September 11-13, Friday to Sunday, for the second, less formal half of our program, consisting of additional lectures, mutual research collaboration, and opportunities for hiking in the mountains.

Contact: Kira Adaricheva or David Stanovsky